

# A Comprehensive Survey on Traffic Missing Data Imputation

Yimei Zhang<sup>ID</sup>, Xiangjie Kong<sup>ID</sup>, *Senior Member, IEEE*, Wenfeng Zhou, Jin Liu<sup>ID</sup>,  
Yanjie Fu, *Senior Member, IEEE*, and Guojiang Shen<sup>ID</sup>

**Abstract**—Intelligent Transportation Systems (ITS) are essential and play a key role in improving road safety, reducing congestion, optimizing traffic flow and facilitating the development of smart cities. The collection of data from ITS and its transformation into information is challenged by the presence of missing data in datasets. Timely and effective handling of missing data is crucial to facilitate intelligent decision making. In response to this need, numerous researchers have proposed various techniques for dealing with missing data, with commendable results in different scenarios. Therefore, this survey aims to provide a well-organized and thorough overview of the research related to imputation of missing data in traffic. Our purposes are at least fourfold. First, we discuss the background of missing data and highlight the value of traffic data imputation. Second, we present a comprehensive list of missing patterns, open data, widely used evaluation metrics and performance goals for this issue. Third, we categorize related studies into three parts: interpolation-based, statistical learning-based, prediction-based methods, and provide a secondary classification within each category to better understand the characteristics and limitations of each method. Finally, we identify future research directions to advance the understanding of traffic data imputation.

**Index Terms**—Traffic data imputation, intelligent transportation systems, deep learning, surveys.

## I. INTRODUCTION

WITH the development of artificial intelligence, intelligent transportation system (ITS) plays an important role in the construction of smart cities. ITS aims to use advanced modern science and technology such as big data mining, cloud computing, and mobile internet to monitor, analyze, and control traffic conditions, so as to improve people's travel efficiency and safety [1].

Traffic data, which contains many types of data such as traffic flow, speed, vehicle trajectory, etc., is one of the foundations of ITS. These data are mainly collected by two



Fig. 1. The process of data acquisition in real world.

types of sensors: one is the fixed sensor, mainly collected by detector equipments such as coil detector; The other is the mobile sensor, mainly based on GPS data, the identification of car license plates, electronic tags to collect data [2]. Despite considerable enhancements in contemporary approaches to gathering information, the transportation network proves to be a complicated system marked by extensive fluctuations and uncertainty [3]. In the actual process of data acquisition, various factors may lead to the interruption and loss of data acquisition: (1) Network interruption or poor communication quality in the process of data transmission. (2) Sensor anomalies caused by natural factors such as construction, road accidents, and weather. (3) In the process of data encryption and implicit processing. It is precisely because of the above factors that missing traffic data has become a practical and common problem (see Fig. 1) in various countries. For example, in the United States, more than 10% of the traffic data collected by the transportation systems of Texas and Georgia are missing, and even the internationally famous traffic flow database PeMS also has more than 5% missing data [4]. Similarly, in Alberta, Canada, more than half of the highway traffic data has missing values, and in extreme cases, the missing value can be as high as 90% [5]. Furthermore, in Beijing, China, the data collected by some detectors has a loss rate of 25%, and the average loss rate of daily traffic volume data is about 10% [6].

In the burgeoning field of intelligent transportation, traffic data imputation is playing an increasingly important role. For example, digital twin technology [7] emerges as an innovative approach offering novel perspectives and solutions for optimizing and planning ITS. However, it needs complete data to achieve dynamic monitoring of roadway infrastructure

Received 23 October 2023; revised 11 March 2024, 4 June 2024, and 31 August 2024; accepted 3 October 2024. Date of publication 23 October 2024; date of current version 27 November 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62072409 and Grant 62073295 and in part by Zhejiang Provincial Natural Science Foundation under Grant LR21F020003. The Associate Editor for this article was T. Tettamanti. (Corresponding author: Xiangjie Kong.)

Yimei Zhang, Xiangjie Kong, Wenfeng Zhou, Jin Liu, and Guojiang Shen are with the College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China (e-mail: yimeizhang0229@gmail.com; xjkong@ieee.org; WenfengZhou98@outlook.com; JinLiu0617@outlook.com; gjshen1975@zjut.edu.cn).

Yanjie Fu is with the School of Computing and AI, Arizona State University, Tempe, AZ 85287 USA (e-mail: yanjie.fu@asu.edu).

Digital Object Identifier 10.1109/TITS.2024.3478816

lifecycles and accurate reconstruction of traffic participants' behavior on road surfaces. Precise navigation is pivotal in the development of safe and reliable autonomous driving systems [8]. Yet, the absence of real-time streaming data can impede the system's perception and decision-making regarding its environment. Additionally, many ITS applications such as real-time adaptive traffic signal control system, predictive bus control framework [9], driver behavior anomaly detection [10] and traffic accident risk prediction system [11], rely on the availability of high-quality data. Research by [12] has shown that the performance of prediction models is significantly influenced by the accuracy of the imputation models employed. Furthermore, the integration of appropriate imputation strategies can enhance both the accuracy and robustness of downstream tasks [13], [14], [15].

Missing traffic data destroys the integrity of the dataset, rendering it incapable of accurately representing traffic information. It not only reduces the accuracy and reliability of traffic conditions, but also affects the decision-making and planning process of the traffic department. Hence, how to deal with missing data becomes one of the hottest topic in traffic field. There are two ways to solve the issue of samples containing missing data: 1) Delete defective samples directly. However, this method will not only result in the loss of some useful information but also may change the data distribution characteristics. Additionally, discrete data will affect the downstream task model to effectively analyze traffic information. In extreme cases, such as a high missing rate, it may also lead to a scene where no data is available in the end. 2) Impute the missing data samples. This kind of method can learn the relationship between data by mining and analyzing the non-missing data information, and infer the missing values to densify the dataset to improve data quality and data availability. Obviously, in most cases, the latter is more reasonable than the former.

Therefore, using data imputation technology to solve the problem of incomplete and low-quality traffic missing datasets can not only provide a high-quality data guarantee for further transportation-related research, but also provide effective data support for transportation authorities to judge the real traffic operation status and propose new control strategies [19]. Besides, using imputation technology can also make data better serve the intelligent transportation system and help the construction of the smart city. This is of great practical significance! In the past few decades, with a lot of research in the fields of traffic prediction, routing, security, etc., researchers have also noticed the importance of the quality of traffic data, and proposed many excellent solutions methods and models. Taking this as the starting point, in this article, we conduct a detailed review of the development of traffic data imputation to provide the most advanced information in this field, which will provide a platform for further research. We also noticed that there are several review works [20], [21], [22], [23] done in the traffic missing data imputation domain in recent years, but they do not overlap significantly with this study. Reference [24] proposed a classic classification of traffic data interpolation methods, namely prediction, interpolation and

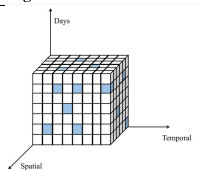
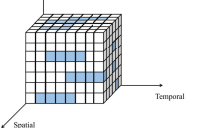
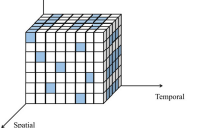
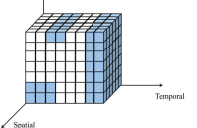
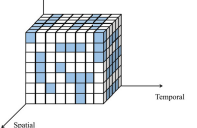
statistical learning methods, and compared the representative methods among them; [23] tested the imputation performance of 13 different methods based on statistical learning for hourly traffic volumes data. While [21] only reviewed the missing data imputation methods from 2006 to 2017, and lacked statistics on deep learning methods. In other words, they lack the latest research progress. However, our paper not only involves statistical analysis of traditional methods but also delves into the latest state-of-the-art research in the field, addressing the gaps left by previous studies. As for [22], they focused their research on traffic temporal data imputation methods which is also their limitation. [25] aimed to review recent papers involving urban networks, where they standardize the terminology used in the classification of missing data and analyze the advantages and disadvantages of different types of methods. This is different from our focus.

Following the above discussions, the main contributions of this study could be summarized as follows:

- 1) Firstly, we provide a comprehensive overview of imputation methods for missing traffic data. From data missing patterns, benchmark datasets, evaluation indicators and performance goals to various types of imputation method models, as well as current challenges and future work in this field. This makes our investigations particularly valuable.
- 2) Secondly, we categorize these existing traffic data imputation methods into three categories: interpolation-based, statistical learning-based, prediction-based methods, according to the current research focus after an in-depth survey. At the same time, a secondary classification is carried out in each category according to the characteristics of the method. Furthermore, we analyze the details and limitations of models.
- 3) Thirdly, we summarize existing publicly available benchmark datasets that are commonly used in the transportation field, including freeway traffic data and urban traffic data, which helps different imputation methods to compare well.
- 4) Finally, we outline possible challenges in the field of traffic data imputation from the perspectives of data missing patterns, joint imputation, signal control effects, model robustness, and real-time performance et al., and discuss some possible future research directions. Our aim is to be able to provide some feasible methods to promote the development of this field.

The rest of the article is arranged as follows: In Section II, we review the classification of data missing patterns in transportation, summarize the public datasets, model performance evaluation metrics, and the goals we pursue in traffic data imputation research. In Section III, the common traffic data imputation methods are divided into three parts. We first introduce the latest advances in interpolation-based methods, and then we focus on discussing the use of statistical learning-based and prediction-based methods for traffic data imputation. Section IV discusses the existing challenges and limitations of traffic data imputation, and identify future research directions. Finally, we conclude the paper in Section V.

TABLE I  
DATA MISSING PATTERNS

| Data Missing Patterns               | Description   | Diagram  |
|-------------------------------------|---|--|
| Missing completely at random (MCAR) | It refers to data loss that is unrelated to any other variables, where the probability of missing data is completely random. The occurrence of missing data is independent of both observed and unobserved data. This type of missingness is characterized by a random scatter plot distribution [16]. MCAR is unpredictable and often caused by factors such as equipment failure, data input errors, or data management mistakes. The advantage of this missing pattern is that it does not introduce bias, and simple imputation methods can be used to fill in the missing values. However, the drawback is that the quantity of missing values can be substantial, and the imputed results may be less accurate due to random errors [16]. |   |
| Missing at random (MAR)             | Time correlated missingness (TCM), where the missing values exhibit time correlation.   |   |
|                                     | Spatially correlated missingness (SCM), where the missing values are correlated with neighboring spatial readings.  |   |
|                                     | Block missingness (BM), where the missing values are correlated both temporally and spatially, forming blocks.  |   |
| Missing not at random (MNAR)        | It means that the absence of data is related to the missing variable itself. In this missing pattern, the missing data cannot simply be estimated from other existing data. Regrettably, most existing missing data techniques struggle to handle MNAR data effectively. It is a crucial and continually developing area of methodological research to create analytical methods for MNAR data. However, these techniques are not yet suitable for widespread use and may have limitations in practical applications [18].  |  |

## II. PRELIMINARIES

Before delving into specific methods, it is essential to understand some preliminary concepts, including the traffic data missing patterns, the public datasets, evaluation metrics for performance assessment, and the goals we pursue in traffic data imputation. These foundational concepts will provide the necessary background and support for the subsequent methods section.

### A. Missing Patterns

The missing datasets formed by different scenarios may have different data characteristics, and the applicable scenarios of different imputation methods are also different. Therefore, understanding the missing patterns of the data is very vital to correctly analyze the data, select the appropriate repair method, and build an accurate model. Through the collation of the existing literature works [17], [90], [91], we find that the statistical missing patterns can be classified into three categories based on the randomness: Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR). These patterns describe relationships between measured variables and the probability of missing data. While these terms have precise probabilistic and mathematical interpretations, they represent three distinct explanations for why data are missing [92]. In this section,

we provide a conceptual overview of each pattern, and additional resources are available for readers seeking more in-depth information on the missing traffic data patterns. The details are shown in the Table I.

### B. Public Datasets

There is a critical problem in the traffic missing data imputation task: the benchmark datasets for comparing model performance. Comparative experiments on benchmark datasets can better demonstrate the performance of the model and make the results more convincing. In realistic intelligent transportation systems, there are various spatio-temporal types of traffic-related data. These data can be divided into data with different attributes from different angles. In this paper, we collect common datasets that are currently used frequently for future researchers to make decisions when conducting experiments. These public datasets are divided into freeway traffic data (**FD**) and urban traffic data (**UD**) from the perspective of road conditions.

The specific information of each dataset is shown in the Table II, which mainly includes the sampling time range, data granularity, spatial coverage, data type, sampling device type and the link address of the dataset. Moreover, we further categorize these datasets into three types [76], namely, large-scale datasets, datasets with recurrence and fluctuation, and datasets with complex spatio-temporal information. Next, we delve

TABLE II  
PUBLIC DATASETS FOR TRAFFIC DATA IMPUTATION

|    | Dataset                              | Time Range             | Granularity | Spatial Coverage          | Data Type                                   | Device | Reference   |
|----|--------------------------------------|------------------------|-------------|---------------------------|---|--------|---|
| FD | METR-LA <sup>1</sup> [26]            | 3/1/2012 - 6/30/2013   | 5 min       | 207 detectors             | Speed                                       | Sensor | [15], [27]–[37]   |
|    | PeMS <sup>2</sup> [38]               | 2001 - 2019            | 5 min       | more than 39000 detectors | Speed / Volume / Occupancy                  | Sensor | [4], [37], [39]–[56]  |
|    | PeMS-4W <sup>3</sup> [57]            | 1/1/2018 - 1/28/2018   | 5 min       | 11160 detectors           | Speed                                       | Sensor | [57], [58]  |
|    | PeMS-8W <sup>3</sup> [57]            | 1/1/2018 - 2/25/2018   | 5 min       | 11160 detectors           | Speed                                       | Sensor | [57]  |
|    | PeMS04 <sup>4</sup> [38]             | 1/1/2018 - 2/28/2018   | 5 min       | 307 detectors             | Speed / Volume / Occupancy                  | Sensor | [33], [59]–[61]   |
|    | PeMS08 <sup>4</sup> [38]             | 7/1/2016 - 8/31/2016   | 5 min       | 170 detectors             | Speed / Volume / Occupancy                  | Sensor | [33], [59]–[61]   |
|    | PeMS-BAY <sup>1</sup> [62]           | 1/1/2017 - 5/31/2017   | 5 min       | 325 detectors             | Speed / Volume / Occupancy                  | Sensor | [28]–[33], [36], [52], [63]–[65]                                      |
| UD | Seattle <sup>5</sup> [66]            | 1/1/2015 - 12/31/2015  | 5 min       | 323 detectors             | Speed                                       | Sensor | [27], [35], [37], [45], [52], [65], [67], [67]–[75]                   |
|    | Portland <sup>4</sup> [70]           | 1/1/2021 - 3/31/2021   | 15 min      | 1156 detectors            | Speed / Volume / Occupancy                  | Sensor | [58], [70], [71], [76], [77]  |
|    | Birmingham parking <sup>7</sup> [67] | 10/4/2016 - 12/19/2016 | 30 min      | 30 parks                  | Car park capacity / Car park occupancy rate | Sensor | [67], [67], [73], [76]  |
|    | Guangzhou <sup>8</sup> [78]          | 8/1/2016 - 9/30/2016   | 10 min      | 214 segments              | Speed                                       | Sensor | [45], [57], [64], [67], [67], [70], [71], [73], [74], [76], [79]–[82] |
|    | TaxiNYC <sup>9</sup> [83]            | 2009 - now             | 60 min      | /                         | Pick-up and drop-off times                  | GPS    | [84]  |
|    | TaxiBJ <sup>10</sup> [85]            | 2013 - 2016            | 30 min      | more than 34000 taxis     | Inflow and outflow                          | GPS    | [86], [87]  |
|    | T-drive <sup>11</sup> [88]           | 2/2/2008 - 2/8/2008    | 177 s       | 10357 taxis               | Trajectory                                  | GPS    | [84], [89]  |

into the characteristics of each category in more detail and discuss their suitability for various types of research, aiding researchers in selecting the most appropriate dataset.

1) *Large-Scale Datasets*: PeMS, PeMS-4W, PeMS-8W, Portland, and Guangzhou datasets are typical large-scale datasets that contain millions of observations with a high number of sensors, making them suitable for testing models' scalability and flexibility in large-scale scenarios. The PeMS, PeMS-4W, PeMS-8W and Portland datasets consist of highway data, while the Guangzhou dataset includes data from urban roads. Therefore, the latter exhibits more fluctuation and randomness.

2) *Datasets With Recurrence and Fluctuation*: METR-LA, PeMS04, PeMS08, PeMS-BAY, Seattle and Birmingham parking dataset are a few of the more commonly used public datasets for traffic data imputation with fluctuation and randomness, while at the same time, appear distinct day-to-day and within-day recurrence [76]. Imputing this kind of data entails capturing details from time series. Of these, METR-LA has more complex data dependencies than other datasets (Los Angeles, which is known for its complicated traffic conditions). The Birmingham parking dataset shows the strong temporal patterns and consistency. In addition, PeMS04, PeMS-BAY and Seattle encompass a greater number of roads and sensors, resulting in a more complex network topology than PeMS08.

3) *Datasets With Complex Spatio-Temporal Information*: TaxiNYC, TaxiBJ and T-drive, collected by mobile sensors, offer higher spatial coverage in urban road networks. The traffic environment of urban is relatively intricate, and road conditions are highly susceptible to random disturbances, resulting in complex spatio-temporal information. These datasets can be used to quantify models' robustness when handling signals with a large variance. Additionally, the TaxiNYC and TaxiBJ datasets include supplementary information such as holidays and meteorological data.

### C. Evaluation Metrics

In the process of traffic data imputation research, it is an important step to select the appropriate model performance evaluation metrics. Different evaluation metrics can show the pros and cons of the model from different view. We summarize some commonly used performance metrics in research, including Mean Absolute Error (MAE), Mean Square Error (MSE), Mean Absolute Percentage Error (MAPE) and their variants. The performance metrics are summarized in Table III above.

### D. Performance Goals

In the previous subsection, the evaluation metrics primarily focus on measuring the discrepancy between the model's estimations and the ground truth. However, relying just on these indicators to judge model performance is inadequate. We must also account for specific scenarios (such as varying missing rates and patterns, real-time applications, large-scale data, and extreme missing situations) to comprehensively evaluate the model's performance from multiple perspectives. Therefore, this subsection will discuss the different goals pursued in traffic data imputation, including accuracy, efficiency, flexibility, robustness, and adaptivity. It is particularly noteworthy that in the subsequent imputation methods part, we will also organize and analyze the literatures on each approach based on these five challenges.

- **Accuracy**: Accuracy refers to how closely the predicted values for missing data align with the ground truth. This is typically measured using various error metrics such as MAE, MSE and MAPE.
- **Efficiency**: Efficiency involves the resources required to handle missing data, such as time, memory, and computational power. This is particularly crucial in real-time applications or when processing large-scale data.
- **Flexibility**: Flexibility refers to models can optimize performance by adjusting parameters according to different



TABLE III  
EVALUATION METRICS IN TRAFFIC DATA IMPUTATION

| Metrics | Description   | Variant               | Formula  |
|---------|---|-----------------------|--|
| MAE     | MAE is the average distance between the predicted values $\hat{x}$ and the true values $\bar{x}$ of a model. Since it directly calculates the average of the residuals, it gives a clear reflection of higher differences. Lower MAE values indicate better imputation accuracy, as the model's predictions are closer to the true values on average.                                 | /                     | $MAE = \frac{1}{n} \sum_{i=1}^n  \bar{x}_i - \hat{x}_i $   |
| MSE     | MSE is a measure of the average squared deviation between the predicted values $\hat{x}$ and the true values $\bar{x}$ . It is calculated by summing the squared differences and dividing by the number of observations $n$ . The MSE provides a non-negative value in the range $[0, +\infty)$ , where a value of 0 indicates a perfect match between the predicted and true values. | RMSE<br>NMSE<br>NRMSE | $MSE = \frac{1}{n} \sum_{i=1}^n (\bar{x}_i - \hat{x}_i)^2$<br>$RMSE = \sqrt{MSE}$<br>$NMSE = \frac{MSE}{\sum_{i=1}^n \bar{x}_i^2}$<br>$NRMSE = \frac{RMSE}{x_{max} - x_{min}}$                   |
| MAPE    | MAPE is a relative measure of error that quantifies the average percentage deviation between predicted values and true values. It is commonly used to compare the accuracy of different time series forecasting models. MAPE avoids the cancellation of positive and negative errors by using absolute values. The smaller the MAPE, the better the quality of the model.             | sMAPE                 | $MAPE = \frac{1}{n} \sum_{i=1}^n \left  \frac{\bar{x}_i - \hat{x}_i}{\bar{x}_i} \right $<br>$sMAPE = \frac{100\%}{n} \sum_{i=1}^n \frac{ \bar{x}_i - \hat{x}_i }{( \bar{x}_i  +  \hat{x}_i )/2}$ |

data characteristics and different missing scenarios, or one single model can handle different data recovery tasks and different missing scenarios.

- **Robustness:** Robustness refers to the ability of a model to provide stable and reliable results when confronted with high missing rates, outliers and noise.
- **Adaptivity:** Adaptivity refers to the capability of a model to exhibit similar performance across different types of datasets or when integrated with downstream tasks.

### III. IMPUTATION METHODS

Across several decades of scholarly exploration in this domain, investigators have presented numerous imputation methodologies grounded in varied techniques to address the issue of missing traffic data. To better understand the development status of this research direction, in this section, we categorize some representative methods into three categories based on the different principles and technical characteristics of each method [44], [93], [94]: interpolation-based methods, statistical learning-based methods, and prediction-based methods. The framework diagram of the classification is shown in Fig. 2. Furthermore, we analyze the application scenarios of various methods, distinguishing between univariate and multivariate approaches (single-sensor vs. multi-sensor). This classification is predicated on the model's capacity to elucidate spatial correlations within traffic data.

#### A. Interpolation-Based Methods

Interpolation-based methods typically utilize the spatio-temporal continuity of the data to estimate missing values from known data points. These methods can be generally divided into three types, namely, temporal-neighboring interpolation, pattern similarity interpolation and spatial neighbor-based interpolation. Temporal-neighboring interpolation methods fill missing points by calculating the observed values from the same detector at the same time on adjacent days or at the same day of neighboring time periods [24]. Pattern similarity

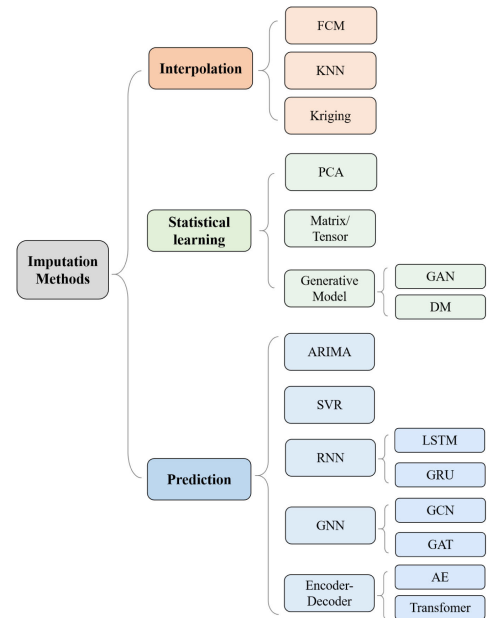


Fig. 2. The organization of imputation methods.

interpolation methods fill missing values by analyzing different patterns in the traffic data and identifying observed points that are most similar to the missing data points based on these patterns. Fuzzy C-Means (FCM) and K-Nearest Neighbors (KNN) are two typical methods in this category. Spatial neighbor-based interpolation is a method for filling missing values by utilizing the spatial correlations among the known data [95] with Kriging being a typical method.

1) *FCM-Based Methods:* FCM classifies the traffic data into different clusters to maximize their similarity by minimizing an objective function to determine the membership degree of each data point and the cluster centroids. Finally, the missing data is estimated through the optimized membership degrees and cluster centroids [96]. Therefore, it is usually

integrated with other optimization algorithms to optimize the above two critical parameters. For instance, Tang et al. [97] employed a hybrid approach combining FCM normalization and Genetic Algorithm (GA) optimization techniques. They converted the traditional vector-based data structure into a matrix-based data structure to better reflect the similarity between different days and the GA is employed to optimize the FCM parameters. In PSO-SVR [98], a hybrid approach integrated FCM with the Particle Swarm Optimization (PSO) algorithm and Support Vector Regression (SVR) to optimize the parameters of FCM for the imputation task. SVR is introduced to build fitness function for FCM optimization and yields more sensible results for outliers, which is robust against the noise. However, its performance is limited in urban arterial roads subject to traffic signal control. FCM is good at utilizing temporal correlation for interpolation, making it suitable for imputing univariate data. However, the efficacy and transferability of FCM in addressing non-recurrent congestion conditions influenced by accidents, extreme weather events, and special circumstances warrant further investigation [97]. Moreover, existing researches on FCM have predominantly focused on imputing missing traffic volume data, other traffic flow variables such as speed and occupancy need to be further explored.

2) *KNN-Based Methods*: The KNN identifies  $k$  samples in the dataset with temporal similarities or spatial similarities through distance measurements, and then uses these samples to estimate the values of the missing data points. In previous research, KNN-based algorithms identify the most similar data by matching each individual link with historical links for each loop detector. The feature matrix generally includes the data from neighboring detectors as follows:

$$X_t^{(s)} = \begin{pmatrix} Z_{t-\tau}^{(s-l)} & \cdots & Z_{t-\tau}^{(s+l)} \\ \vdots & \ddots & \vdots \\ Z_{t+\tau}^{(s-l)} & \cdots & Z_{t+\tau}^{(s+l)} \end{pmatrix}, \quad (1)$$

where  $Z_t^{(s)}$  is the observed value at time intervals  $t = \{1, 2, \dots, T\}$  and spatial locations  $s = \{1, 2, \dots, S\}$ . The variable  $l$  represents the spatial range of the neighboring detectors in the KNN search, and  $\tau$  represents the temporal range of neighboring time intervals. This strategy of searching individual links requires a high computation power. Therefore, Tak et al. [99] and Cai et al. [39] improved the efficiency of matching with reasonable accuracy. The former divided road sections based on the correlation coefficient with a moving window between detectors, while the latter enhanced computational performance by reducing data size and using localized data for calculations. However, both methods lack robustness when dealing with high missing data rates due to their reliance on information provided by neighbors. Sun et al. [40] alleviated this issue by extending the neighbor count using a window approach. KNN accomplishes missing value filling by daily similarity and road proximity of traffic data, making it applicable to both univariate and multivariate. This algorithm is relatively simple and has lower computational overhead [100]. Therefore, it performs well with small data volumes and simpler scenarios. However, this method

encounters hardships in large-scale datasets and when the traffic state is more complex [101]. Furthermore, in case of high missing rates, the lack of similar data will seriously affect the interpolation accuracy [102]. The performance of this method depends on the parameter selection, and the prediction of  $k$  is a difficult problem, increasing the complexity in practical applications [99].

3) *Kriging-Based Methods*: Kriging estimates the time series of new locations by taking advantage of the spatial correlations among the observed locations [105]. There are two primary types of kriging that are commonly used in traffic missing data repair: Kriging-based and Cokriging-based [95]. Kriging encompasses Simple Kriging (SK) and Ordinary Kriging (OK), and they can be expressed as follows:

$$\hat{X}(s_0) = m + \sum_{\alpha=1}^n w_{\alpha}(X(s_{\alpha}) - m), \quad (2)$$

$$\hat{X}(s_0) = \sum_{\alpha=1}^n w_{\alpha}X(s_{\alpha}), \quad (3)$$

where  $s_0$  is a vector of the location where  $\hat{X}$  will be predicted,  $m$  is a mean of dataset,  $w_{\alpha}$  is a weight of an observation  $\alpha$ ,  $s_{\alpha}$  is a vector of the location where an observed  $X$  is placed on a spatio-temporal plane, and  $n$  is the number of observations used for imputation. Cokriging is inherently the multivariate extension of kriging [106]. It allows to use secondary data sources to complement observed primary data. Similar to Kriging, there are two types of Cokriging: Ordinary Cokriging (OCK) and Simple Cokriging (SCK), which can be represented as follows:

$$\hat{X}_{i_0}(s_0) = \sum_{i=1}^q \sum_{\alpha=1}^{n_i} w_{\alpha}^i X_i(s_{\alpha}), \quad (4)$$

$$\hat{X}_{i_0}(s_0) = m_{i_0} + \sum_{i=1}^q \sum_{\alpha=1}^{n_i} w_{\alpha}^i [X_i(s_{\alpha}) - m_i], \quad (5)$$

where  $\hat{X}_{i_0}(s_0)$  is estimated value of a primary variable  $i_0$  at an unobserved location vector  $s_0$ ,  $q$  is the number of variables, and  $n_i$  is the number of observed locations of  $i$  th variable.  $m_{i_0}$  is the global mean of a primary variable, and  $m_i$  is the global mean of  $i$  th variable [95]. Bae et al. [95] proposed two cokriging methods that take advantage of the spatio-temporal dependency in traffic data and multiple data sources to impute high-resolution traffic speed data. However, it has limitations in accurately interpolating average annual daily traffic under complex traffic patterns caused by different road functions or land uses. To address this problem, Ma et al. [104] proposed a copula model that combines spatial dependence and marginal distributions. In recent years, the problem of estimating the traffic state of undiscovered road connections has been the subject of many studies. This issue is often framed as spatio-temporal traffic kriging [107]. Lei et al. [27] proposed a Bayesian Kernel Matrix Factorization (BKMF) model to leverage the data from available sensors to estimate/interpolate variables at unseen locations. BKMF enables flexible hyper-parameter tuning for spatio-temporal kriging and exhibits reduced sensitivity to increasing missing rates. However, it suffers from high consumption cost. The inference process for kernel hyperparameters and latent factors takes time and memory when the data size is large. To scale up the kriging on large-scale network, Nie et al. [58] sped up

TABLE IV  
SUMMARY OF INTERPOLATION-BASED METHODS ON TRAFFIC DATA IMPUTATION

| Reference        | Data type                  | Technique   | Evaluation metric | Missing pattern          | Missing rate |
|------------------|----------------------------|-------------|-------------------|--------------------------|--------------|
| FCMGA [97]       | Volume                     | FCM         | RMSE, r, PAE, RA  | MAR                      | 1% - 30%     |
| PSO-SVR [98]     | Volume                     | FCM, SVR    | RA, PAE, RMSE     | MCAR, MAR                | 1% - 30%     |
| KNN-DI [99]      | Link travel time           | KNN         | MAPE, RMSE, PCV   | MCAR, TCM, SCM, BM       | 0% - 30%     |
| GSW-kNN [40]     | Flow                       | KNN         | RMSE              | MCAR, MAR                | 5% - 90%     |
| ST-KNN [39]      | Volume                     | KNN         | MAE, RMSE, MAPE   | MCAR                     | 20%          |
| CKDI [95]        | Speed                      | Kriging     | MAE               | MCAR, TCM, SCM, BM, MNAR | 10% - 40%    |
| ST-Kriging [103] | Volume / Speed / Occupancy | Kriging     | MAD, RMSE         | MCAR                     | 1.0% - 36.1% |
| CopulaDI [104]   | Volume                     | Kriging     | MSE, MAPE         | MCAR, MAR                | 25% - 75%    |
| BKMF [27]        | Speed                      | MF, Kriging | MAE, RMSE         | MCAR                     | 20% - 95%    |
| LETC [58]        | Speed                      | TC, Kriging | MAE, RMSE         | MCAR                     | 20% - 70%    |

their model by randomized tensor singular value decomposition and conjugate gradient method, which achieves both accuracy and computational efficiency. The Kriging method is able to take full advantage of the spatio-temporal correlation of traffic data, so it is suitable for multivariate data. The advantages of this method include its ability to recover data for road sections without any observations, addressing the Newell's third-detector problem [108], and the construction of spatial correlations is not limited to using geographical similarity but includes fundamental diagram relationships as well [58]. It demonstrates that sparse sensors can achieve a "collaborative perception" of traffic state for undetected road links by effectively leveraging their inherent multi-dimensional correlations [58]. However, most existing methods describe the relationships between measurement points using a simple distance metric or function to model the spatial correlations between sensors [109], which works well on highways, while their performance in dealing with more complex urban networks needs further exploration.

In order to make readers understand the introduced interpolation-based methods more intuitively, we summarize them from the following aspects based on the original classification: model name, data type, technique, evaluation metric, missing pattern and missing rate. We present these results in the form of a table, hoping to help researchers conduct follow-up research. For specific information, see Table IV.

### B. Statistical Learning-Based Methods

Statistical learning-based methods hypothesize the existence of simplifiable underlying characteristics (data distributions or low-rank structures) within traffic data, leveraging these characteristics to fill in missing values [110]. Approaches such as Principal Component Analysis (PCA), Matrix/Tensor and Generative Models utilize the above features to impute traffic data.

1) *PCA-Based Methods*: PPCA [111] is one of the most representative methods based on the PCA approach. It considers the primary portion of the observed data (here, traffic flow time series) as a linear mapping of a collection of Gaussian distributed data [112], as shown in Fig. 3. Distributing a priori knowledge offers valuable insights into the dynamics of traffic flows beyond mere similarity, thereby enhancing the

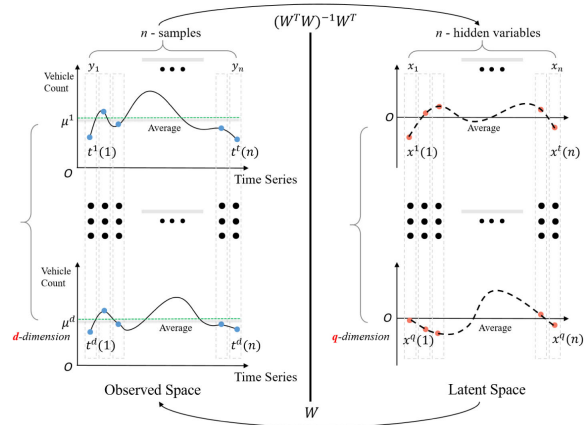


Fig. 3. An illustration of PPCA model [111].

performance of traffic flow estimation. However, in PPCA, the observed data has a linear relationship with the latent variables, which does not satisfy the nonlinear characteristics of traffic data. Subsequently, Kernel Probabilistic Principal Component Analysis (KPPCA) and Mixed Probability Principal Component Analysis (MPPCA), as variants of PPCA, are also used for traffic data imputation. References [4] and [43] show that, KPPCA and MPPCA are able to capture nonlinear spatio-temporal dependencies in traffic data compared to PPCA, and therefore obtain higher imputation accuracy. As mentioned above, whether PCA-based method is suitable for processing univariate or multivariate data depends on the specific implementation of the method. PCA-based methods reveal key patterns and trends in traffic flow data, making them effective for traffic data imputation. Furthermore, studies demonstrate its applicability in large-scale and real-time scenarios [113]. Robust principal component analysis can easily detect anomalies or abnormal patterns in traffic data when a component significantly increases covariance [114]. When the missing data rate is not high, its good computational efficiency and considerable imputation accuracy make it a favorable choice. As the missing rate increases, the main information in the data is corrupted leading to its performance degradation. In addition, it is challenging to generalize preconceived assumptions to other cases due to the heterogeneity of traffic spatio-temporal data [45].

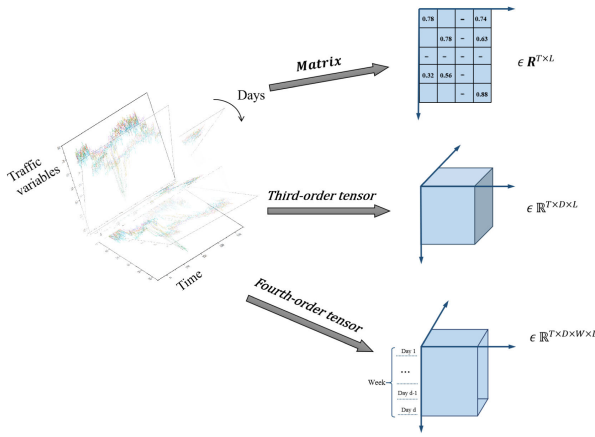


Fig. 4. Three manifestations of traffic data.

2) *Matrix/Tensor-Based Methods*: For traffic data, it can be represented in different data forms such as matrix (*road segment*  $\times$  *time interval*), third-order tensor (*road segment*  $\times$  *day*  $\times$  *time interval*), fourth-order tensor (*road segment*  $\times$  *week*  $\times$  *day of week*  $\times$  *time interval*) [79]. An example diagram of them is shown in the Fig. 4. Tensor structures outperform matrix representations in preserving information within datasets which have large-scale historical traffic data spanning months or years, as tensors more effectively capture temporal correlations (such as day-of-week patterns and periodic trends). In contrast, matrix representations are better suited for real-time applications involving smaller datasets. Matrix/tensor-based methods are able to fully utilize spatial and temporal information in transportation data to capture underlying patterns and relationships. Through multi-dimensional operations and decomposition, they can effectively handle complex spatio-temporal data. Thus, these techniques are excellent for interpolating data that is multivariate.

a) *Matrix-based*: Matrix-based methods can be divided into two categories: Matrix Factorization (MF) and Matrix Completion (MC). MF uses matrix factorization models (Singular Value Decomposition (SVD) and Non-negative Matrix Factorization (NMF), etc.) to decompose the original matrix into the product of two or more low-rank matrices. The potential spatio-temporal features embedded in these low-rank matrices are utilized for traffic data restoration. MC involves using MF or other optimization algorithms to find matrices that best fit the known data distribution patterns, thereby filling in missing values. Temporal and Adaptive Spatial constrained Low Rank (TAS-LR) [115] and Hessian Regularization Spatio-Temporal Low Rank (HRST-LR) [71] decomposed traffic data into two matrices: a temporal feature matrix which describes time-varying characteristics, and a spatial feature matrix which describes the nature of road segments, and then applied constraints on the feature matrix to impute missing data. HRST-LR captures the temporal evolution of traffic data via a second-order difference of time-series constraint presenting better stability than TAS-LR at a high missing data rate. TAS-LR constructs an adaptive affinity matrix with Laplacian regularization to capture spatial neighborhood

similarity, while HRST-LR solves the spatial similarity by introducing Hessian Energy. Laplacian regularization does not effectively preserve the locality of the data space. In contrast, HRST-LR maintains the locality of the data structure well by incorporating Hessian Energy, leading to higher imputation accuracy in SCM scenario. Nonetheless, HRST-LR can be computationally less efficient than TAS-LR in certain cases. Matrix-based methods can simultaneously consider the temporal and spatial correlation of traffic data, utilizing the intrinsic correlations within existing data for effective imputation. However, traffic data has remarkable multi-dimension correlations, e.g., day-to-day recurrence, neighborhood similarity, within-day regularity [76], and the two-dimensional matrix representation is no longer able to cover enough traffic information, resulting in limited performance.

b) *Tensor-based*: To compensate for the lack of matrix, Tan et al. [5] first introduced the matrix extension of tensor model for modeling traffic data. One problem for tensor is to identify the most appropriate traffic data representation (third-order tensor or fourth-order tensor) for missing data imputation tasks. Chen et al. [79] showed that the fourth-order tensor does not show better imputation results than the third-order tensor, but rather increases the computational cost. Therefore, subsequent studies tend to use a third-order tensor to represent traffic data. These methods can be broadly classified into two categories, Tensor Decomposition (TD) and Tensor Completion (TC) [116].

Compared to MF, TD is a multilinear structure that considers multiple aspects of the data. Currently, there are two main classical formulas for TD, the first is the Tucker decomposition, which extracts the core tensor as an instrument to capture intrinsic correlations in the data, which is usually presented as:

$$M_{i,j,k} = \sum_{m=1}^{r_1} \sum_{n=1}^{r_2} \sum_{l=1}^{r_3} (g_{m,n,l} u_{i,m} \circ v_{j,n} \circ w_{k,l}), \quad (6)$$

where  $\mathcal{G}^{r_1 \times r_2 \times r_3}$  is the core tensor,  $g_{m,n,l} \in \mathcal{G}^{r_1 \times r_2 \times r_3}$ , and  $u_{i,m} \in \mathbb{R}^{I_1 \times r_1}$ ,  $v_{j,n} \in \mathbb{R}^{I_2 \times r_2}$ , and  $w_{k,l} \in \mathbb{R}^{I_3 \times r_3}$ , with ranks expressed as  $r_1, r_2, r_3$  [5], [73]. Tucker decomposition offers the advantage of retaining more information and enhancing model flexibility, but they may also result in an increased number of parameters, thereby raising computational overhead [73]. The second is the CP decomposition, which expresses a higher-order tensor as a sum of rank-one tensors, and always denoted as:

$$M_{i,j,k} = \sum_{r=1}^R \lambda_r a_r \circ b_r \circ c_r, \quad (7)$$

where  $\circ$  means the outer product,  $R$  stands for latent rank which is a non-negative integer,  $\lambda_r \in \text{diag}(\Lambda)$ ,  $a_r \in \mathbb{R}^{I_1}$ ,  $b_r \in \mathbb{R}^{I_2}$ , and  $c_r \in \mathbb{R}^{I_3}$  for  $r \in \{1, \dots, R\}$  [73], [79]. CP decomposition offers the advantage of reducing the number of parameters and simplifying model complexity, but it may also result in information loss and overfitting [73]. To overcome this problem, [79], [82], [117] incorporated the Bayesian approach into CP decomposition to achieve stable performance. Multi-Task Neural Tensor Factorization (MTNTF) method [118] addressed the joint imputation of traffic speed



and volume by employing a neural network to replace CP factorization to learn the non-linearities between latent factors in multiple data imputation tasks. Xing et al. [119] integrated homogeneous data fusion into the CP tensor decomposition framework for dealing with the long-term missing in interval-wise data imputation.

Unlike the use of a pre-defined rank in the TD methods, the idea of TC is to directly minimize the tensor rank. To tackle rank minimization, most methods utilize the tensor Nuclear Norm (NN) as a convex surrogate for the minimization [76]. However, the problem of over-relaxation in NN makes them not achieve the desired performance in practice [76]. In recent years, researchers began to explore and design new nonconvex rank functions to substitute for NN. For example, LRTC-TNN [67] employed the Truncated Nuclear Norm (TNN) for tensors to impute traffic data and introduced a truncation rate parameter to globally control the numbers of contributing singular values. However, the truncated singular values of TNN still face the issue of over-relaxation, which can hinder the method from accurately capturing the real low-rank structure. To solve this problem, Nie et al. [76] proposed a truncated tensor Schatten  $p$ -norm (TSpN) that jointly absorbs the merits of both TNN and Schatten  $p$ -norm. The proposed TSpN stands for a better tensor rank surrogate with more accuracy and flexibility. Improving imputation accuracy in datasets characterized by complex time series signals heavily relies on capturing the smoothness and dynamics of temporal evolution. Low-rank Autoregressive Tensor Completion (LATC) [70] achieved this by integrating temporal modeling with an autoregressive process and Robust Tucker factorization-based Tensor Completion (RTTC) [81] employed a time series decomposition model to account for outliers and temporal pattern shifts. To enhance model flexibility, [73] employed a Sigmoid mapper to release non-negative constraints during training and [64] introduced an implicit regularizer to exploit the nonlocal self-similarity prior of traffic data, which can boost the imputation performance. In addition, to improve the model's flexibility to accommodate the heterogeneity of traffic states and multiple corrupted scenarios, Hu et al. [45] proposed SCPN based on the tensor Schatten capped  $p$  norm, which effectively balances the rank and NN. In recent decades, the proliferation of sensors has made big data a major trend in the real world, posing a challenge for large-scale traffic data interpolation. Low-Tubal-Rank Smoothing Tensor Completion (LSTC-Tubal) [57] and Laplacian Enhanced low-rank Tensor Completion (LETTC) [58] converted the tensor completion problem into a series of small-scale subproblems and then connected their solutions by linear transformations, which reduces the computational cost and allows the tensor-based approaches to scale to large-scale data.

Matrix/Tensor-based models are able to fully exploit the spatio-temporal intrinsic structure in traffic data and provide useful priors [45]. Because of their potent spatio-temporal modeling capabilities, when faced with extreme scenarios and intricate missing data patterns, they can maintain relatively stable performance [33]. Zhang et al. [14] point out that matrix/tensor-based interpolation models are computationally expensive. Liang et al. [30] also argue that the model should

be inductive to meet the real-time requirement, which means no retraining when new data arrives. While most matrix/tensor completion methods are transductive and unable to generalize to the next time-window. These studies seem to suggest that matrix/tensor methods are less applicable in real-time scenarios. However, the framework designed by Yang et al. [37] demonstrates that with appropriate optimization and real-time adaptation techniques, matrix/tensor-based methods can be applied in online scenarios. In addition, for the large-scale traffic data characterized by the dimensions of "location", "time interval of day", and "day", the tensor-based approaches are more suitable as they achieve more accurate results by incorporating the richer temporal relationships present in historical data [57], [58]. By leveraging linear transforms, it becomes technically feasible to address large-scale tensor problems by decomposing them into a series of smaller subproblems. However, during the process of learning low-rank representations, these models may excessively smooth the reconstructed data, consequently filtering out informative signals and leading to an over-squashing reconstruction in certain scenarios [33]. In addition, if the data do not meet the low-rank assumption, matrix/tensor-based methods will struggle to effectively reconstruct missing values.

3) *Generative Models*: Generative models generate new samples by learning the underlying distribution of the data. Their primary objective is to emulate the data generation process, enabling the creation of new samples that resemble the training data, accomplishing missing data imputation task. Their flexible structure facilitates seamless integration with other techniques, making them particularly suitable for interpolating multivariate data. The main generative models are Generative Adversarial Network (GAN) and Diffusion Model (DM).

a) *GAN-based*: GAN is used as a generative model to generate new samples based on the trained distribution by training two networks against one another. The standard GAN consists of two parts, the generator and the discriminator, whose structure is shown on the left in Fig. 5, and it achieves high-quality data generation through a zero-sum game with each other. On the basis of standard GAN, Yoon et al. [120] introduced the hint mechanism into the discriminator, and proposed a missing data imputation model GAIN (as shown on the right in Fig. 5) which realized the fine-grained true and false discrimination of data. Different from the former, Kazemi and Meidani [121] proposed an iterative GAN model named IGANI. In this approach, the generator iterates twice to generate estimated data, subsequently used to train a robust discriminator. This discriminator can identify the primary imputed data as real and the secondary data as false, thereby assisting in improving the generator's ability. However, the above-mentioned approaches overlook the spatio-temporal correlations among the traffic data. Therefore, most of the subsequent methods model the spatio-temporal representation in traffic data through the network design of generators and discriminators. For example, Spatio-Temporal GAN (STGAN) model [122] used dilated convolutions in the generator to capture the spatio-temporal dependency and establishes an

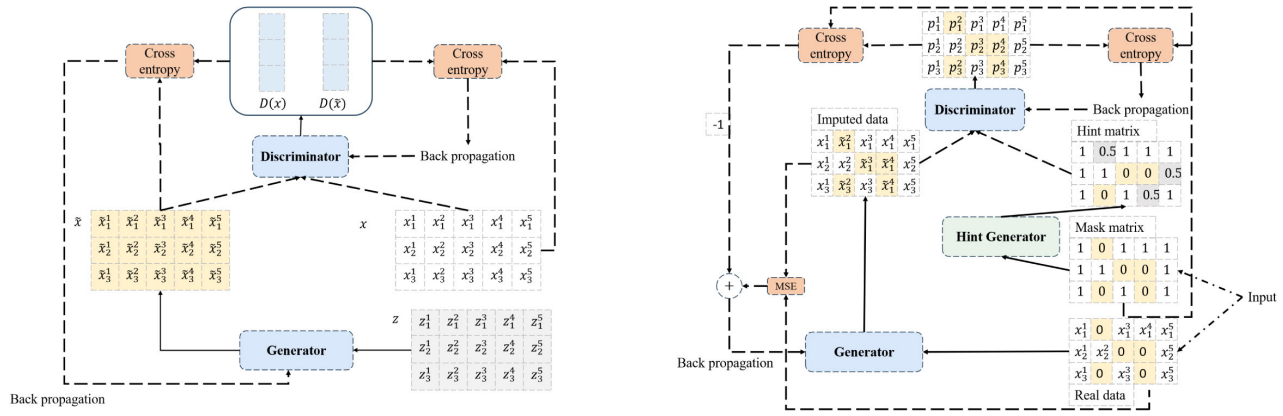


Fig. 5. The structure of GAN and GAIN [120].

index to record missing entries in the incomplete matrix, guiding the generator network by focusing on the missing point. Considering the advantages of attention mechanism to choose from the excess of information that is available, Self-Attention Generative Adversarial Imputation Net (SA-GAIN) [69] employed a self-attention mechanism to capture the correlation between spatially distributed sensors at different time points. Spatio-Temporal Learnable Bidirectional Attention Generative Adversarial Networks (ST-LBAGAN) [84] utilized a learnable bidirectional attention graph and set the input as a multi-channel matrix composed of temporally related data, thereby effectively capturing the spatio-temporal stochastic characteristics in traffic flow. Furthermore, Gao et al. [123] found that the traditional loss functions of GAN may not be suitable due to the unique nature of traffic data. Therefore, the use of a custom loss function promotes a better adaptation of the model to scenarios with missing traffic data. The most commonly used method is to constrain the difference between the predicted and the true values through reconstruction loss [41], [84], [124], [125]. Yang et al. [84] introduced perceptual loss and style loss to capture the higher-level semantic information in the data set. Considering the strong correlation between traffic data and its surrounding neighbors, STGAN [122] adopted center loss to ensure that the estimated entries conform to the distribution of their corresponding neighbors. The aforementioned methods all focus on imputing missing data at the granularity of road segment and aggregated time intervals, Gated attentional GAN (GaGAN) [94] imputed missing data at the lane and signal cycle level and introduced an attention-based generator to capture the spatio-temporal correlations in traffic sequence data constrained by signal timing. In the face of large-scale road networks and high missing data rates in the real world, some methods may struggle with efficiency and accuracy. Han et al. [44] proposed a batch-oriented data imputation method and incomplete traffic tensors were recovered by back-propagation algorithms based on the trained GAN. Zhang et al. [46] proposed dynamic multi-level generative adversarial networks to effectively complete the traffic data imputation task of large-scale multi-sensors distributed on the road network. In addition to the above, [59], [68], [72] have also attempted to use GAN to accomplish the task of traffic data imputation and have achieved promising results. GAN learns in detail using

an adversarial approach from observed data, producing highly accurate results [100]. It is less sensitive to missing rates and, through the design of loss functions and discriminators, can robustly handle high missing rate problems and complex data missing patterns [100]. Moreover, with appropriate training strategy and network design, it can be applied to large-scale scenarios. However, its instability during training and high resource demands make it challenging in resource-constrained environments.

b) *DM-based*: As a rising star in probabilistic generative models, diffusion model has received great attention in the field of deep learning. Compared to GAN, diffusion-based imputation methods offer a high degree of flexibility in assuming real data distributions and have a more stable training process. They recover original data from diffuse (noisy) data by gradually inverting the diffusion process. CSDI [128] first introduced DM into data imputation tasks. However, it simply learns the spatio-temporal dependence of the observation part, leading to biased predictions of noisy. Additionally, the noise prediction module in CSDI employs two Transformers, and the risk of memory overflow increases when the time series is long. To address the first problem, Liu et al. [28] proposed a conditional diffusion framework for spatio-temporal imputation with enhanced prior modeling, named PriSTI. It integrates the interpolated conditional information and geographic information as a condition for denoising, which makes it achieve excellent imputation results in traffic data. To address the second problem with CSDI, SSSD [29] replaced Transformers with structured state space model to solve the quadratic complexity issue. Recently researchers [63], [129], [130] also adopted new noise prediction methods to explore the potential of DM in the field of data imputation. However, most of the existing models are designed specifically for multivariate time series or spatio-temporal data. The development of models specifically tailored to the characteristics of traffic data to accommodate different traffic demands is necessary. DM emerges as a powerful generative tool with a strong capacity for capturing intricate traffic data patterns. However, they are still computationally complex, which poses a challenge for their use in resource-constrained or real-time environments [131]. Furthermore, they have problems in dealing with boundary coherence between missing and observed parts [132].

TABLE V  
SUMMARY OF STATISTICAL LEARNING-BASED METHODS ON TRAFFIC DATA IMPUTATION

| Reference         | Data type                         | Technique      | Evaluation metric    | Missing pattern | Missing rate |
|-------------------|-----------------------------------|----------------|----------------------|-----------------|--------------|
| BPFA [126]        | Volume                            | PCA            | p-RMSE, p-MAED       | MCAR, MAR, MNAR | 4% - 46%     |
| PPCA [111]        | Volume                            | PCA            | NMAE, NRMSE, RMSE    | MCAR, MAR       | 0% - 50%     |
| KPPCA [4]         | Volume                            | PCA            | NMAE, NRMSE, RMSE    | MCAR, MAR       | 5% - 30%     |
| MPPCA [43]        | Volume                            | PCA            | RMSE                 | MCAR, MAR       | 5% - 40%     |
| TAS-LR [115]      | Speed / Volume                    | MC             | NMAE, RMSE           | MCAR, TCM, SCM  | 0% - 100%    |
| AL-SRMF [42]      | Speed / Volume                    | MC             | NMAE, RMSE           | MCAR            | 20% - 90%    |
| HRST-LR [71]      | Speed / Volume / Occupancy        | MC             | MAPE, RMSE, MAE      | MCAR, TCM, SCM  | 0% - 100%    |
| BATF [82]         | Speed                             | TD             | MAPE, RMSE           | MCAR, MAR       | 10% - 50%    |
| BGCP [79]         | Speed                             | TD             | MAPE, RMSE           | MCAR, MNAR      | 10% - 50%    |
| DFCP [119]        | Volume                            | TD             | MAPE, RMSE           | MCAR, MNAR      | 10% - 50%    |
| MTNTF [118]       | Speed / Volume                    | TD             | MAPE, MAE, RMSE      | MCAR, MAR       | 10% - 90%    |
| LTRR-NLS [64]     | Speed / Flow                      | TD             | MAE, MRE, MAPE, RMSE | MCAR, TCM       | 50% - 90%    |
| FNNTL [80]        | speed                             | TD             | MAPE, RMSE, MAD      | MCAR, MAR       | 10% - 90%    |
| TMac-TT [127]     | Volume                            | TC             | RSE                  | MCAR            | 10% - 90%    |
| LRTC-TNN [67]     | Flow / Speed / Occupancy          | TC             | MAPE                 | MCAR, MAR       | 10% - 70%    |
| LATC [70]         | Speed / Volume / Flow / Occupancy | TC             | MAPE, RMSE           | MCAR, MNAR, BM  | 20% - 40%    |
| LSTC-Tubal [57]   | Speed                             | TC             | MAPE, RMSE           | MCAR, MAR       | 30% - 70%    |
| LRTC-TSpN [76]    | Flow / Speed / Volume / Occupancy | TC             | MAE, RMSE            | MCAR, MAR       | 10% - 95%    |
| SCPN [45]         | Speed / Volume                    | TC             | RMSE, MAPE           | MCAR, MNAR      | 20% - 80%    |
| RTTC [81]         | Speed                             | TC             | MAE, SMAPE           | MCAR, TCM, BM   | 10% - 96%    |
| NT-DPTC [73]      | Flow / Speed / Occupancy          | TC             | MAE, NMAE, RMSE      | MCAR            | 10% - 95%    |
| CLRTR [74]        | Speed                             | TC             | MAPE, RMSE           | MCAR, MAR       | 10% - 95%    |
| IIVMTV [75]       | Speed / Volume / Occupancy        | TC             | MAPE, RMSE           | MCAR, MAR, MNAR | 20% - 80%    |
| TNN-HTV [56]      | Speed                             | TC             | MAPE, RMSE           | MCAR, MAR       | 50% - 90%    |
| STGAN [122]       | Speed                             | GAN, AE        | NMAE, RMSE, MAE      | MCAR, TCM, SCM  | 20% - 80%    |
| CA-GAN [44]       | Speed                             | GAN            | MRE                  | MCAR, TCM       | 20% - 80%    |
| ST-LBAGAN [84]    | Trajectory                        | GAN, AE        | RMSE, MAE            | MCAR, TCM, SCM  | 10% - 60%    |
| SA-GAIN [69]      | Volume                            | GAN, AE        | MAE, MMD, RMSE       | BM, TCM, SCM    | 10% - 80%    |
| PD-GAN [41]       | Volume                            | GAN            | MAE, RMSE, MRE       | MCAR            | 30% - 80%    |
| MST-GAN [59]      | Volume                            | GAN, LSTM, GCN | RMSE, MAE            | RM, TCM, SCM    | 30% - 80%    |
| GAE-GAN-LSTM [72] | Speed                             | AE, GAN, LSTM  | RMSE, MAE, MAPE      | MCAR            | 10% - 90%    |
| TGAIN [68]        | Volume / Speed                    | GAN            | RMSE, MAPE           | MCAR, MAR       | 10% - 90%    |
| GaGAN [94]        | Speed                             | GAN, GRU, GCN  | MAE, RMSE, MAPE      | MCAR, MAR, MNAR | 20% - 100%   |
| ST-DIGAN [125]    | Volume                            | GAN, GCN       | MAE, RMSE, MAPE      | MCAR, MAR       | 10% - 70%    |
| MLGAN [46]        | Volume                            | GAN, GCN, GRU  | MAE, RMSE, $R^2$     | MCAR            | 5% - 95%     |
| CSDI [128]        | Occupancy                         | DM             | MAE, MSE, CRPS       | MCAR, BM        | 10% - 90%    |
| PriSTI [28]       | Speed                             | DM             | MAE, MSE             | MCAR, BM        | 10% - 90%    |
| SSSD [29]         | Speed                             | DM             | MAE, RMSE, CRPS      | MCAR            | 25%          |
| MIDM [63]         | Volume                            | DM             | MAE, MSE, MRE        | MCAR, BM        | 10% - 90%    |

Generative models reveal complex patterns and dynamic characteristics in traffic data by learning implicit representations. Furthermore, these models demonstrate high accuracy when handling large volumes of traffic data and more accuracy is obtained even when the missing rate increased [100]. However, they require high-performance computing resources. Additionally, the opacity and lack of interpretability within the models present obstacles for researchers seeking to improve and optimize them [133].

As we did in subsection III-A, we summarize the representative statistical learning-based methods in a table for reader to better understand. The fields are the same as in Table IV. For specific information, see the Table V.

### C. Prediction-Based Methods

Prediction-based methods estimate missing data by using historical observations. The missing values are considered as

predicted values. Auto Regressive Integrated Moving Average (ARIMA), Support Vector Regression (SVR), Recurrent Neural Network (RNN), Graph Neural Network (GNN), and Encoder-Decoder Architecture (EDA) utilize the above features to impute traffic data.

1) *ARIMA-Based Methods*: ARIMA can automatically identify time-correlated changes in traffic data, and use existing data to predict and fill missing values [134]. This model involves estimating three key parameters:  $p$  for auto-regression component,  $d$  for integration component, and  $q$  for moving average component. The general equation for the ARIMA model is as follows:

$$X_t = c + \sum_{i=1}^p \alpha_i x_{d_{n-i}} + \sum_{i=0}^q \beta_i e_{n-i}, \quad (8)$$

where  $x_d$  is the observations  $X$  differenced by  $d$  times and  $e$  is the estimation error.  $\alpha_i$  and  $\beta_i$  are the parameters of the model, which are used to express the relationship between the current value and the value of the past  $p$  time points and the

relationship between the current value and the error of the past  $q$  time points respectively,  $c$  is a constant term. Elshenawy et al. [135] proposed an imputation model based on ARIMA to estimate missing highway traffic volume data automatically. Their approach involves a systematic mechanism for determining the required ARIMA parameters. In addition, several extended ARIMA models have been studied, such as seasonal ARIMA [136], [137], fractionally integrated vector autoregressive and moving average (VARFIMA) [138]. ARIMA relies on historical data to capture seasonality and trends within a time series, rendering it suitable for forecasting applications characterized by short durations [135], [139]. Therefore, this method is good at dealing with univariate data. However, traditional AIRMA ignores the spatial correlation between sensors thus it does have limitations when it comes to capturing complex spatio-temporal relationships in traffic datasets. This gave rise to the Space-Time Autoregressive Integrated Moving Average (STARIMA). Recently, [77] achieved online traffic prediction with missing data by combining STARIMA and PPCA. Traffic flow usually shows a significant temporal pattern which can be effectively captured using ARIMA. However, these models are typically linear and may not fully capture all the complexities present in traffic data.

2) *SVR-Based Methods*: SVR predicts missing values by using historical data to construct regression models that capture non-linear relationships between data. In [48], SVR is used to directly predict missing values. While in [140], it is used to evaluate the accuracy of estimated values instead of directly estimating. Moreover, Luo et al. [141] combined SVR with deep belief network for traffic flow data imputation. SVR uses a kernel function to deal with non-linear relationships in data. Therefore, the parameters determination for the nonlinear kernel function is important to data imputation. For example, in [48] and [142], researchers used intelligent optimization algorithms such as GA and PSO algorithm to find the optimal estimate. SVR possesses strong nonlinear modeling capabilities and is able to capture complex spatial relationships in traffic data, which makes it suitable for imputation of multivariate data [143]. Advantages of this approach are the efficient in massive dimensional areas and efficient memory consumption. When confronted with nonlinear relationships and complex missing data patterns, SVR offers a powerful and robust solution. However, it often needs to be combined with other methods to capture complex spatio-temporal dependencies in traffic data.

3) *RNN-Based Methods*: There are multiple reasons for fluctuation in traffic data, and it is difficult for ARIMA and SVR to adequately learn the complex information involved. The currently emerging deep neural networks, on the other hand, can provide more sophisticated data imputation strategies for traffic data due to large number of parameters. As we know, traffic data can be represented as a type of data collected over time and can be considered essentially as a time series data. RNN [144] is a memory-capable neural network that effectively captures time-related characters in sequence data. The main feature of RNN is to capture contextual information and dependencies in data by iteratively updating hidden states, which gives it a natural advantage to predict missing data using

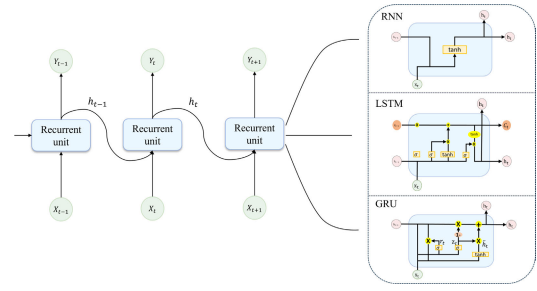


Fig. 6. The internal recurrent neuron structures of RNN, LSTM and GRU.

historical observations. The state update process of RNN is as follows:

$$\begin{aligned} h_t &= f(W \cdot h_{t-1} + U \cdot x_t + b), \\ o_t &= g(V \cdot h_t + c), \end{aligned} \quad (9)$$

where the hidden state  $h_t$  is determined by the input data at the current moment and the hidden state  $h_{t-1}$  at the previous moment,  $o_t$  is the output state,  $W, U, V, b, c$  are learnable parameters,  $f, g$  are the activation functions. RNNs have an advantage in dealing with variable-length sequences due to their recursive formulation, which makes them particularly effective for predicting historical and future data values from a single sensor. But for mining spatial location relationships in the data, other techniques and models are required. There are two popular variants of RNN: Long Short-Term Memory (LSTM) [145] and Gated Recurrent Unit (GRU) [146], which are proposed to solve the short-term memory problem of RNN due to gradient disappearance. The folding form of the three recurrent neurons is shown in Fig. 6.

a) *LSTM-based*: As a variant of RNN, its core idea is to regulate the information flow through a gate mechanism to effectively capture the long-term dependencies in sequence data. Inspired by the varying contributions of information at different time steps to inference, Tian et al. [147] proposed LSTM-M, which utilizes a multi-scale temporal smoothing method to impute missing values. However, this method can only use the data before the missing value for interpolation. Based on the fact that time series data can be inferred from the forward direction as well as the backward direction. Cao et al. [148] proposed a novel bi-direction imputation model BRITS for multivariate time series data. Similarly, [35], [149], [150], [151] designed two LSTM networks, one in the forward time direction and the other in the backward direction to exploit information from both the past and the future. Yang et al. [37] proposed LSTM-GL-ReTF for real-time data repair and prediction. When the data missing rate is high, bidirectional LSTM serves to further stabilize the model. In order to handle missing data without relying on any predefined value, [65] designed an “imputation unit” in LSTM structure, which can automatically generate interpolated values during the loop. The proposed architecture, possibly containing multiple layers of LSTM or BDLSTM components, can be flexible for solving different tasks (e.g. imputation and prediction). However, due to stacking multiple layers of networks, the imputation efficiency of this model is affected. Therefore, [152] used



a multilayered LSTM network with a parameter transfer strategy. Moreover, to overcome the limitation of LSTM in modeling the dependence of long time series data, [153] introduced the attention mechanism to improve it. The LSTM-based model is good at processing sequence information, so it can fully capture the temporal information in traffic data. However, the acquisition of spatial information often needs to be completed in conjunction with other methods.

b) *GRU-based*: As another variant of RNN, GRU also utilizes gate mechanism to control the information flow. Its advantage is that it has fewer parameters, but can achieve similar effects to LSTM. So it seems to be more popular in the application process. Luo et al. [154] first modified the GRU structure and designed a novel GRU variant GRUI, which can consider the non-fixed time lag and dilute the influence of past observations determined by the time lag. At the same time, [155] introduced a GRU-D model, which can effectively exploit two different forms of missing information: data missing pattern and sampling interval. However, when dealing with traffic data, the disadvantage of the pure GRU-based method is also obvious, it cannot mine the spatial dependencies of the data. Therefore, many researchers have combined GRU with other deep learning methods [14], [60], [156], aiming to overcome this limitation. Among them, [60] utilized an auxiliary GRU to model the missing patterns of the missing data, allowing the model to extract more valuable information from various types of data. GRU is similar to LSTM and is good at capturing the temporal correlation of traffic data. It has fewer parameters and is usually more efficient compared to LSTM.

RNNs are adept at capturing complex nonlinear features and dynamically changing patterns in traffic data through their recurrent structure, enabling effective learning of temporal recurrences and variations. Furthermore, they can utilize the data before and after the missing points for forward and backward imputation to provide more information around the missing point. However, there are situations where employing bidirectional context could result in data leakage, such as when utilizing the imputed data as input for a forecasting task in downstream applications [29]. Yang et al. [37] also demonstrate that while bidirectional RNN models increase stability when the rate of missing data is high, they may result in overly smoothed predictions when the missing rate is low. Additionally, RNNs are prone to gradient vanishing or explosion when dealing with long sequences [157], and their sequential nature hinders the parallelization of the training process [158]. It is also found in [58] that the RNN-based approaches are not suitable for network-wide applications due to the significant training costs and memory consumption. Therefore, their efficiency and scalability are limited when handling large-scale data.

4) *GNN-Based Methods*: RNN has demonstrated remarkable performance in temporal modeling. However, traffic road networks exhibit an intricate graph structure, leading to complex spatial and temporal dependencies in traffic data. GNN is widely used for modeling and analyzing graph-structured data due to their proficiency in handling non-Euclidean data. And, this also makes it currently considered as one of the most

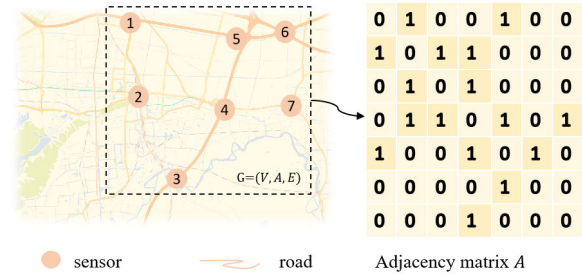


Fig. 7. The process of adjacency matrix construction.

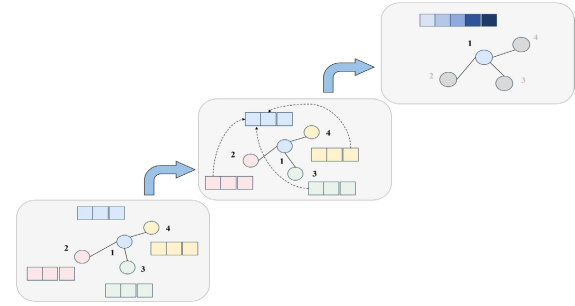


Fig. 8. The calculation process of GCN [165].

advanced and effective techniques for processing traffic spatio-temporal data. Based on the adjacency relationship of nodes, GNN updates the state of nodes by cyclically exchanging neighborhood information until a stable equilibrium is reached, which can be summarized as [159]:

$$h_i^{(t)} = \sum_{j \in N(i)} f(X_i, X_{(i,j)}^e, X_j, h_j^{(t-1)}) \quad (10)$$

where  $N(i)$  is the set of neighboring nodes of node  $i$ ,  $f(\cdot)$  is a function that integrates the information from node  $i$  and its neighbor  $j$ ,  $X_i$  and  $X_j$  represent the initial features of node  $i$  and  $j$ ,  $X_{(i,j)}^e$  represents the features of the edge between nodes  $i$  and  $j$ , and  $h_j^{(t-1)}$  is the state representation of node  $j$  at the previous iteration  $t-1$ . As mentioned, node adjacency is crucial for GNN. In the field of traffic, researchers usually define the traffic network as a weighted or unweighted, directed or undirected graph  $G = (V, E, A)$  based on node connectivity. Fig. 7 provides an intuitive adjacency matrix construction process of unweighted graph. As a neural network that acts directly on graph structures, GNN is able to extract complex non-linear correlations between irregular road networks through graph structure modeling [160], [161], [162], [163] and make inferences based on the data represented by the graphs, making it well suited for handling spatial information and interpolating multivariate data. During the development of GNN, two important variants have evolved: graph convolutional network (GCN) and graph attention network (GAT). We next will discuss the application of these two methods to traffic data imputation.

a) *GCN-based*: Considering that the spatial dependencies of traffic data are influenced by direction within the traffic network, [30], [35], [149] employed a diffusion graph convolutional network [62] to model traffic flow as a diffusion process and capture the stochastic spatial dependencies in various

directions. At the same time, under the environment of missing data, the construction of dynamic road network graph to mine the implicit spatial correlation of road network to achieve data restoration has also received attention [32], [35], [60]. They learned the dynamic spatial dependencies of network structures by generating trainable dynamic neighbor matrices. Traditional geographically defined adjacency matrices may not be able to contain full spatial correlation, e.g., two nodes have similar traffic states but are far away from each other. The studies utilized attention mechanisms [31], adaptive adjacency matrices [15] and an external attention-enhanced diffusion convolution [30] to capture the global semantic similarity. To meet the real-time requirement, GCBRNN [14] designed a graph convolution gated recurrent unit which has a fast inference speed, and Spatial-Temporal Aware data Recovery network (STAR) [30] got the imputed data by induction without retraining the whole model. However, GCNs only consider the local neighbor nodes, the contribution weights from the local neighbor nodes to the central node are explicitly predefined [164]. In addition, they fail to estimate traffic state on undetected road links, so additional techniques or methods may be needed to deal with unseen graph structures.

*b) GAT-based:* In real-world transportation networks, the graph is large-scale and combined with noise, the traffic status also changes continuously over time. Efficiently extracting useful features from these graphs is difficult. One solution is to incorporate attention mechanisms into graph networks, namely GAT. GAT uses the attention mechanism to dynamically calculate fusion weights based on the characteristics of adjacent nodes, which enables it to effectively capture the dynamic spatial correlation within data according to different road structures. The calculation process of GAT as shown in Fig. 9. Ye et al. [124] first applied GAT to the traffic data imputation task to model dynamic influence. Similarly, the studies [34], [52], [61], [166] generated spatial representation of the missing data by considering the relationships between neighboring nodes, which helps in understanding localized patterns and road segment interactions essential for accurate imputation. Unlike previous work that imputed traffic speed or flow, Zhao et al. [167] proposed an imputation framework based on map-embedded for vehicle trajectories. The framework utilized a distance-based social attention mechanism information aggregation with the angle-based social attention mechanism effectively emphasized key neighbor relationships, and achieved excellent trajectory repair task. Compared to GCN, GAT introduces an attention mechanism that adaptively assigns importance weights to adjacent nodes. This allows GAT to more flexibly handle the heterogeneity and diversity in traffic data, providing stronger adaptivity. Furthermore, it is directly applicable to inductive learning problems, including tasks where the model has to generalize to completely unseen graphs [168]. One limitation of GAT is that it typically only considers one-hop neighboring nodes, neglecting the entire graph structure [164].

The transportation system naturally presents a graph structure, enabling GNNs to effectively capture the complex spatial dependencies within the traffic network. Furthermore, many studies have also demonstrated the feasibility of GNNs in

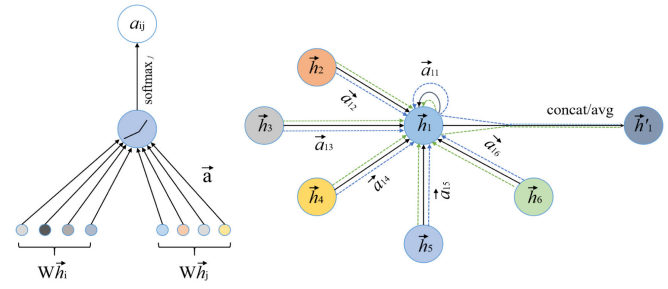


Fig. 9. The calculation process of GAT [168].

real-time scenarios, as they can perform inference rapidly without requiring model retraining. Although traditional GNNs may have limitations in capturing long-term spatial dependencies [164], improvements such as incorporating the temporal dimension and designing dynamic graph structures can effectively enhance their ability to capture long-term spatial dependencies in traffic data [62], [169]. However, GNN-based methods rely too much on supervised learning, which may lead to insufficient robustness and generalization capabilities in the face of high missing rates and noisy data. Furthermore, GNNs present significant challenges on large-scale traffic data with pretty large “locations” dimension. Firstly, due to the propagation rules of GNNs, which require information transmission and aggregation across the entire traffic network, the computational resource demands increase substantially as the size of the network grows, with an increase in the number of nodes (e.g. sensors) and edges (e.g., roads or connections) [165]. Secondly, the issue of over-smoothing becomes particularly prominent in large-scale graphs, potentially leading to node features becoming similar and thus affecting the model’s performance [170], [171]. Although some optimization methods, such as sampling [172] and graph partitioning [173], [174] can alleviate these issues to a certain extent, effectively scaling GNNs to large-scale traffic networks while maintaining model performance remains a challenge in practical applications [58].

*5) Encoder-Decoder Architecture:* RNNs focus on temporal information, whereas GNNs focus on spatial information; however, the distinguishing feature of traffic data is its spatio-temporal correlation. Therefore, simultaneous integration of temporal and spatial information is necessary. EDA contains two main parts: the encoder and the decoder. The encoder is responsible for extracting the feature representation of the input. The decoder, in turn, receives the output of the encoder and generates predictions. Due to the flexibility of the architecture, EDA can integrate technologies such as CNN, RNN and GNN to enable it to process spatio-temporal traffic data more efficiently. For example, Zhuang et al. [177] developed a CNN-based context encoder to transform raw traffic data into 2D images to reconstruct the complete image from the missing source. However, this approach, utilizing conventional CNN layers focusing on independent road segments, falls short in modeling non-Euclidean spatial correlations in complex traffic networks. Ye et al. [124] introduced a graph attention convolutional network following an encoder-decoder structure for end-to-end traffic data imputation. Furthermore, some

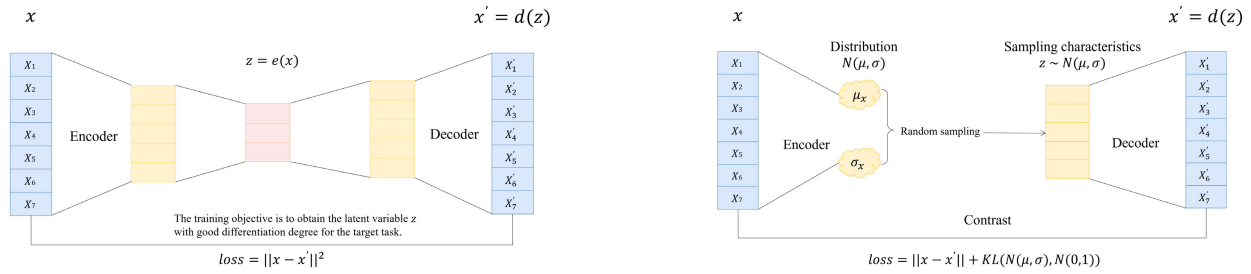


Fig. 10. The structure of Autoencoder [175] and Variational Autoencoder [176].

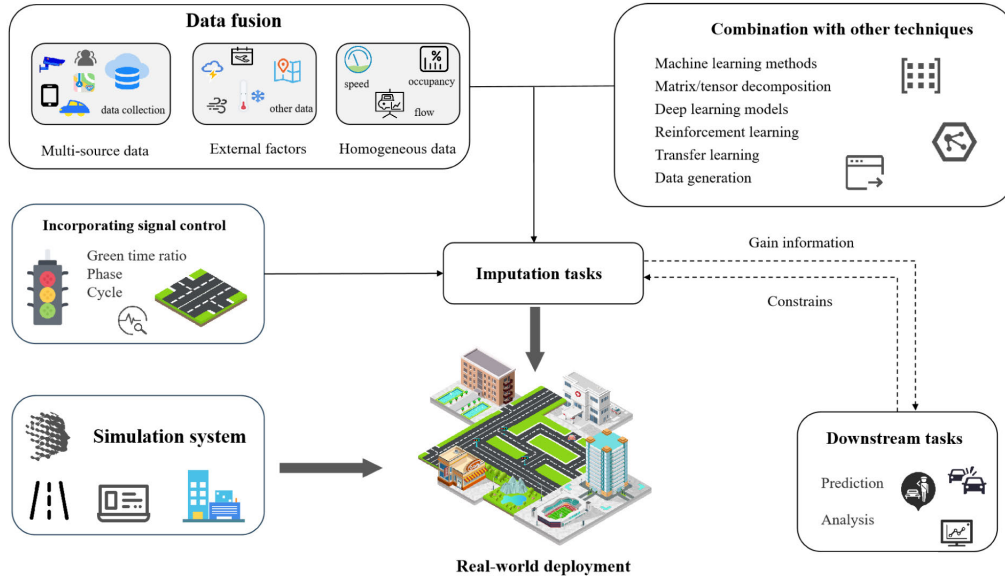


Fig. 11. Future work on traffic data imputation.

studies [15], [178], [179] adopted EDA to jointly optimize data imputation and downstream tasks to improve the performance and generalization capabilities of the model. EDA can dynamically adapt to traffic data in different scenarios through appropriate modifications and the introduction of more complex structures, so it is suitable for dealing with both univariate and multivariate data. There are two important types of members in EDA: AutoEncoder (AE) and Transformer.

*a) AE-based:* In AE, the encoder is responsible for data dimensionality reduction and representation learning, and the decoder is responsible for mapping the output of the encoder to the original data space. The structure of AE is shown on the left in Fig. 10. Duan et al. [55], [180] first applied AE in traffic data imputation. They proposed a Denoising Stacked AutoEncoders (DSAE) structured with a DAE at the bottom and stacked AEs in the middle layers. However, DSAE regards each day's traffic flow data as a vector and barely leverages spatial information, which is suitable to repair an isolated traffic sensor or a road segment [35]. Ye et al. [181] proposed an ensemble AE to solve this problem. Variational AutoEncoder (VAE) combines the ideas of autoencoder and variational reasoning to further enhance the data generation ability of AE. The structure of VAE is shown on the right in Fig. 10. [54] and [182] used VAE to capture the distribution and underlying characteristics of traffic data. [183] enhanced VAE by Gaussian mixture distribution to improve

the generalization and accuracy of the imputed data. However, due to the entanglement of potential variables in a VAE, its interpretability can be poor. There have been some studies devoted to learning to represent entanglement [184], [185], [186], which shows that the well-disentangled representation can improve the performance and robustness of the algorithm. AE-based imputation model is robust and compact in the final encoding layer, which can be applied to many types of traffic data. However, overfitting can occur due to a lack of data on using the overparameterized model.

*b) Transformer-based:* Transformer employs EDA, which addresses the limitations of RNNs (i.e., parallel computation is not possible when dealing with serial data). [47] is the first to study the problem of interpolating missing traffic data using a Transformer-based deep learning framework. The applied self-attention mechanism can model the spatial-temporal relationship among traffic data in a more reasonable way, making the model more robust even under high missing rate conditions. Reference [49] accomplished the interpolation of missing data by identifying the causal network of observations and then combining it with a spatio-temporal Transformer. Furthermore, [53] leveraged Graph Transformer and Large Language Model (LLM) to solve large-scale traffic data imputation problems. Due to the lack of prior knowledge about the latent spatio-temporal processes in traffic data, Transformer may mistakenly retain

TABLE VI  
SUMMARY OF PREDICTION-BASED METHODS ON TRAFFIC DATA IMPUTATION

| Reference         | Data type                  | Technique            | Evaluation metric | Missing pattern    | Missing rate |
|-------------------|----------------------------|----------------------|-------------------|--------------------|--------------|
| ARIMA-DI [135]    | Volume                     | ARIMA                | MAE, MAPE, RMSE   | MCAR               | 0.5% - 72.6% |
| PPCA-STARIMA [77] | Speed / Volume / Occupancy | ARIMA, PPCA          | MAPE, RMSE        | MAR, MNAR          | 10% - 50%    |
| MVLM [48]         | Volume                     | LSTM, SVR            | MAPE, PV, r       | MCAR, SCM, TCM, BM | 5% - 50%     |
| PSO-SVR [98]      | Volume                     | FCM, SVR             | RA, PAE, RMSE     | MCAR, MAR          | 1% - 30%     |
| SBU-LSTM [65]     | Speed                      | LSTM                 | MAE, MAPE, RMSE   | MAR, MNAR          | 10% - 80%    |
| M-LSTM [152]      | Speed                      | LSTM                 | RMSE              | /                  | /            |
| LSTM-GL-ReMF [37] | Volume / Speed / Occupancy | LSTM, MF             | MAPE, RMSE        | MCAR, SCM          | 10% - 40%    |
| GCRINT [149]      | Volume                     | LSTM, GCN            | MAE               | MCAR, RM           | 50% - 90%    |
| MGIA [150]        | Volume                     | LSTM, GCN            | RMSE, MAE, MAPE   | TCM, MCAR          | 10% - 60%    |
| GCBRRN [14]       | Speed / Volume             | GRU, GCN             | MAE, MAPE         | MAR                | 20% - 70%    |
| STAR [30]         | Speed                      | GNN                  | MAE, RMSE, MAPE   | MCAR, MAR, MNAR    | 20%          |
| DGCRIN [60]       | Speed                      | GRU, GCN             | MAPE, RMSE, MAE   | MCAR, TCM, SCM     | 30% - 70%    |
| GRIN [36]         | /                          | GNN, AE, GRU         | MAE, MSE, MRE     | BM, MCAR           | 5% - 25%     |
| MACRO [151]       | Volume                     | GCN, LSTM            | MAE, RMSE         | MCAR               | 10% - 50%    |
| MDGCN [35]        | Speed                      | GCN, LSTM            | MAE, RMSE, MAPE   | MCAR, TCM, SCM, BM | 20% - 80%    |
| ST-GIN [34]       | Volume                     | GAT, GRU             | MAE, MSE          | MAR, MNAR          | 10% - 70%    |
| SAGCRN [31]       | Speed                      | GCN                  | RMSE, MAE, NMAE   | MCAR, TCM, SCM, BM | 10% - 60%    |
| RIHGCN [50]       | Speed                      | GCN, LSTM            | MAE, RMSE         | MCAR               | 20% - 80%    |
| RDGCNI [51]       | Speed / Volume             | GCN                  | MAE, MAPE, RMSE   | MCAR, BM           | 30% - 70%    |
| ADGCN [32]        | Flow                       | GCN                  | MAE, MSE, MAPE    | MCAR, BM           | 25% - 75%    |
| HSTGCN [52]       | Speed                      | GCN                  | MAE, RMSE, MAPE   | MCAR, TCM, SCM     | 10% - 70%    |
| GSTAE [15]        | Speed                      | GCN, GRU             | MAE, MAPE, RMSE   | MCAR, TCM          | 20% - 80%    |
| TVI-GNN [166]     | Volume                     | GAT, LSTM            | MAPE, RMSE        | MAR                | 10% - 80%    |
| GACN [124]        | Speed                      | GAT, EDA             | MAPE              | MCAR, BM           | 10% - 90%    |
| GARNN [61]        | Volume                     | LSTM, GAT            | MAPE, RMSE, MAE   | MCAR, TCM, SCM     | 10% - 50%    |
| TrajGAT [167]     | Trajectory                 | GAT, AE, LSTM        | ADE, FDE          | MCAR, MAR          | 0% - 50%     |
| CNNED [177]       | Volume                     | EDA                  | RMSE, MAE, MPE    | MCAR               | 5% - 50%     |
| DSAE [55]         | Volume                     | AE                   | MAE, RMSE, MRE    | MCAR, MAR          | 5% - 50%     |
| AE-Ensemble [181] | Speed                      | AE                   | MAPE              | MCAR, TCM, SCM, BM | 10% - 50%    |
| VAERL [54]        | Speed                      | VAE                  | RMSE, MAPE        | MCAR, MNAR         | 10% - 40%    |
| TINet [47]        | Speed                      | Transformer          | MAE, RMSE, MAPE   | MCAR, BM           | 10% - 90%    |
| GT-TDI [53]       | Volume / Speed             | GNN, Transformer     | MAAPE, RMSE       | MCAR, MNAR         | 10% - 90%    |
| CDSTN [49]        | Speed                      | Transformer          | MAPE, RMSE        | MCAR               | 30% - 70%    |
| ImputeFormer [33] | Speed / Volume             | Transformer, TC, GNN | MAE               | MCAR, MAR          | 25% - 95%    |

high-frequency noise in the data as informative signals, leading to overfitting. Therefore, Nie et al. [33] proposed a low rankness-induced Transformer model to address this issue. By combining the advantages of low-rank structural priors and the expressive features of Transformers, the proposed method demonstrates high computational efficiency, generalizability across various datasets, versatility for different scenarios, and interpretability of results. The Transformer model, through its flexible token input multi-head attention mechanism, dynamically captures rich implicit patterns in spatio-temporal data, demonstrating excellent accuracy and robustness [19]. In addition, Transformer addresses the limitations of RNNs in handling long-range dependencies, parallelization challenges, gradient vanishing, and model flexibility through key innovations such as self-attention mechanisms, layer normalization, and residual connections. Although the large number of parameters and quadratic increase in computational complexity make Transformers limited on resource-constrained hardware [47], the patch

design of vision Transformer offers new insights for efficiency improvements [187], [188].

EDA is particularly notable for its efficient generative output and high expressivity, allowing more effective imputations in most scenarios. Due to its flexible architecture, it can capture the spatio-temporal correlations in traffic data by combining the respective advantages of RNNs and GNNs. In addition, its natural encoder and decoder architecture makes it particularly suitable for multi-task scenarios. However, the high model complexity and the need for substantial amounts of data and computational resources for training present a challenge for their application in resource-constrained or real-time environments.

As we did in subsection III-A and III-B, we summarize the representative prediction-based methods in a table for reader to better understand. For specific information, see the following Table VI. Additionally, for some more challenging scenarios such as extreme data missing (more than 90% missing), large-scale data, and real-time data imputation, we have summarized



TABLE VII  
SUMMARY OF LITERATURES IN CHALLENGING SCENARIOS

| Method            | Accuracy | Efficiency | Flexibility | Robustness | Adaptivity | Scenario    | Task  |
|-------------------|----------|------------|-------------|------------|------------|-------------|---|
| GSW-kNN [40]      | ✓        |            |             | ✓          |            | EMS         | Volume  |
| TAS-LR [115]      | ✓        |            |             | ✓          | ✓          | EMS         | Speed, volume, OD flow                              |
| TMac-TT [127]     | ✓        |            |             | ✓          |            | EMS         | Volume  |
| AL-SRMF [42]      | ✓        |            |             | ✓          |            | EMS         | Speed   |
| HRST-LR [71]      | ✓        |            |             | ✓          | ✓          | EMS         | Speed, volume, occupancy                            |
| MTNTF [118]       | ✓        |            |             | ✓          | ✓          | EMS         | Speed, volume                                       |
| LATC [70]         | ✓        |            |             | ✓          | ✓          | EMS         | Speed, volume, metro passenger flow                 |
| LRTC-TNN [67]     | ✓        |            | ✓           | ✓          | ✓          | EMS         | Speed, metro passenger flow, park occupancy         |
| NT-DPTC [73]      | ✓        |            | ✓           | ✓          | ✓          | EMS         | Speed, metro passenger flow, park occupancy         |
| CLRTR [74]        | ✓        |            | ✓           | ✓          | ✓          | EMS         | Speed   |
| GaGAN [94]        | ✓        |            | ✓           | ✓          |            | EMS         | Lane-level speed                                    |
| CSDI [128]        | ✓        |            |             | ✓          |            | EMS         | Occupancy   |
| PriSTI [28]       | ✓        |            |             | ✓          |            | EMS         | Speed   |
| MIDM [63]         | ✓        |            |             | ✓          |            | EMS         | Volume  |
| GACN [124]        | ✓        |            |             | ✓          |            | EMS         | Speed   |
| TINet [47]        | ✓        |            |             | ✓          |            | EMS         | Speed   |
| RTTC [81]         | ✓        |            |             | ✓          | ✓          | EMS         | Speed, volume                                       |
| BKMF [27]         | ✓        |            | ✓           | ✓          |            | EMS         | Speed   |
| ImputeFormer [33] | ✓        | ✓          | ✓           | ✓          | ✓          | EMS         | Speed, volume                                       |
| LSTC-Tubal [57]   | ✓        | ✓          |             | ✓          |            | Large       | Speed   |
| CA-GAN [44]       | ✓        | ✓          |             | ✓          |            | Large       | Speed   |
| LETC [58]         | ✓        | ✓          | ✓           | ✓          |            | Large       | Speed   |
| LSTM-GL-ReMF [37] | ✓        | ✓          |             |            | ✓          | Real-time   | Speed, volume, occupancy; Imputation and prediction |
| GCBRNN [14]       | ✓        | ✓          |             | ✓          | ✓          | Real-time   | Speed, volume; Imputation and prediction            |
| PPCA-STARIMA [77] | ✓        | ✓          |             |            | ✓          | Real-time   | Speed, volume, occupancy; Imputation and prediction |
| STAR [30]         | ✓        | ✓          | ✓           | ✓          | ✓          | Real-time   | Speed; Imputation and prediction                    |
| TrajGAT [167]     | ✓        | ✓          |             |            |            | Real-time   | Lane-level trajectory                               |
| TVI-GNN [166]     | ✓        | ✓          |             | ✓          |            | Real-time   | Volume  |
| FNNTL [80]        | ✓        | ✓          |             | ✓          |            | EMS / Large | Speed   |
| GT-TDI [53]       | ✓        |            |             | ✓          | ✓          | EMS / Large | Speed, volume                                       |
| MLGAN [46]        | ✓        |            |             | ✓          |            | EMS / Large | Volume  |
| SCPN [45]         | ✓        |            | ✓           | ✓          | ✓          | EMS / Large | Speed, volume                                       |
| LRTC-TSpN [76]    | ✓        | ✓          | ✓           | ✓          | ✓          | EMS / Large | Volume, OD flow, speed, occupancy                   |

\* EMS: Extreme Missing Scenario, Large: Large-scale Scenario, Real-time: Real-time imputation Scenario.

the literatures from the perspective of five performance goals (i.e., accuracy, efficiency, flexibility, robustness, adaptivity) in Table VII. These scenarios are typically more challenging and require further exploration in future studies.

#### IV. CHALLENGES AND FUTURE WORK

This section will highlight some of the most pressing problems yet unresolved in the area of traffic data imputation and suggest some potential future research routes.

##### A. Challenges

Although the imputation research of traffic missing data has made considerable strides and the approaches stated above have been able to improve imputing efficiency, there are still some important issues that need to be addressed. We summarize several pivotal challenges from previous research and present some of our own ideas. Compared to others, we think

the following challenges are more relevant and practically feasible. In future research, researchers should consider how to address these issues. The main open challenges are summarized below:

- 1) **Missing data patterns:** As discussed earlier, wrecks such as device collection and storage failures, natural disasters, and privacy issues resulting in the deletion of certain data points inevitably corrupt data collection. These various causes give rise to distinct patterns of missing data [30], [33]. Different missing patterns impart unique characteristics to datasets, allowing the imputation process to choose suitable models based on these features, thereby enhancing overall performance. Therefore, understanding and modeling these missing data patterns is one of the challenges in traffic data imputation.
- 2) **Joint imputation:** Traffic datasets obtained from sensors often contain multiple types of data such as traffic

volume, speed, occupancy, and more, with implicit correlation features between these traffic variables (which may be useful for traffic data imputation) [109], [119]. Furthermore, the process of data collection and transmission can also lead to different types of datasets exhibiting different missing rates and missing patterns, and the inconsistency between the two causes difficulties in data modelling [122]. Moreover, if the data imputation process only focuses on a certain feature, it may result in the imputation results not accurately fitting the real values because of the lack of additional information to guide the process. Most current imputation models only consider a single type of data at a time, ignoring the consideration of such implicit correlations between multiple features.

- 3) **Signal control effects:** One difference between urban traffic and freeway traffic is that their data patterns are heavily influenced by traffic signal factors. It is not difficult to find that the traffic characteristics of intersections controlled by traffic signals clearly have a coupling relationship with the signal control, and also have an impact on their surrounding intersections. Simply applying freeway traffic data imputation model to urban traffic data often yields poor results because the modelling of the role of traffic signals is neglected [94]. Therefore, the modelling of traffic signal control is a non-negligible factor, as well as a difficult one, when performing imputation on urban traffic data.
- 4) **Robustness:** Uncertainties in data types, missing rates, and missing patterns lead to unstable model performance [30], [33]. Most of the existing models may perform better in common scenarios (unique data type, low missing rate, and single missing mode). But when dealing with other types of data or in some extreme cases (high missing rate, mixed missing mode), the model performance is not satisfactory. In the real world, the distribution and rate of missing data are uncontrollable, and currently available models are helpless in these extreme cases. Furthermore, core parameters selection and potential anomalies in observations will also affect the robustness of the model [45].
- 5) **Real-time and online:** Real-time processing is constrained by limited computational resources and strict time limits for producing solutions. However, many existing imputation methods with excellent results (e.g., deep learning models) usually require a large amount of computing resources due to high computational complexity, resulting in too long calculation time, which makes it impossible to fill missing values online [14]. Furthermore, with the growth of communication devices and the Internet of Things [189], the amount of digital data we create is growing exponentially [110]. As massive data collection becomes the norm, the size of incomplete datasets continues to grow, making it more challenging to train imputation models to real time. This requires that the imputation technology needs to be low-complexity and high-performance in order to meet the actual requirements. Hence, how to

deploy data imputation technology to real life for online missing values estimation is a big challenge at present.

## B. Future Work

As modern technology continues to advance, we envision that future research on traffic data imputation can hone in on the following key aspects:

- 1) **Data fusion:** As data sources for ITS increase, future research can explore how to integrate multiple data sources, including traffic sensor data, mobile device data, and social media data. At the same time, joint imputation between different traffic data variables can also be considered, such as integrating volume, speed, occupancy, etc. Taking advantage of the multi-source and multi-variable nature of traffic data, we can further mine the spatio-temporal characteristics of traffic and improve the quality of data imputation [109], [119]. In addition, auxiliary data such as weather information and POI can also be incorporated, which can significantly affect traffic patterns and can provide additional gain information for the task [60].
- 2) **Combination of multiple techniques:** Since traffic data has multiple attributes and missing patterns, it can be combined according to the pros of different methods to maximize the advantages of the technique. An ensemble learning approach can be adopted to select appropriate methods and design corresponding modules, while maintaining accuracy, robustness, and efficiency [51], [80], [100]. In addition, data-driven models (e.g., deep learning methods) may not perform well in complex missing scenarios where there is insufficient training data, which can employ self-supervised learning or supplementing the data with generative models.
- 3) **Parallel with downstream tasks:** Most of the existing imputation methods are usually used as a data preprocessing step for downstream tasks. In future work, the multi-task model of them can be considered. Logically, imputation and downstream tasks are a complementary relationship. The interpolation task can provide gain information for the downstream task, and the downstream task constrains the imputation task in turn [15]. In addition, the multi-task model can also save the cost of model design and calculation.
- 4) **Incorporating signal control effects:** In urban traffic, future research can explore methods that explicitly account for the influence of signal control on traffic situation. This could involve developing models that consider signal timings, phasing, and coordination to improve the imputation of missing values in signalized road environments. By incorporating signal control effects, imputation methods can better capture the intricacies of traffic patterns and facilitate more accurate urban traffic analysis and prediction.
- 5) **Robustness improvement:** The complexity and uncertainties inherent in real-world scenarios pose significant challenges for models, highlighting the importance of enhancing their robustness. For instance, incorporating ensemble learning and data augmentation methods

can improve the stability of the model. Research suggests [109] that the distribution of masking samples of training data and missing observations of testing data should be close to ensure good performance. Therefore, adopting appropriate masking strategies can improve model robustness, such as training the model with a hybrid masking strategy [33]. Moreover, LLM, as a new modeling paradigm, demonstrates robust predictive performance even when dealing with limited datasets [190]. It is characterized by strong generalization capabilities, data efficiency, multimodal knowledge, and easy optimization [191]. Therefore, the integration of LLM holds great promise as a forward-looking direction to improve model's robustness.

- 6) **Real-time performance:** Online data imputation based on real-time data streams is essential for the ITS, especially in situations requiring rapid decision-making and response. Therefore, improving the computational efficiency of models and reducing resource consumption to meet the accuracy and quick response needs is an important research direction [100]. One feasible approach is to integrate the imputation model with downstream tasks, leveraging complementary features and parameter sharing to save model computation time [37]. In addition, it is interesting to further improve the computational efficiency of the models with the help of advanced computational frameworks, such as GPU platforms or quantum computing, or the use of distributed computation and parallel processing techniques, which distribute computational tasks to multiple devices or machines for collaborative processing.
- 7) **Practical applications in real-world ITS scenarios:** According to our survey results, traffic data imputation primarily focuses on the simulations with historical traffic data. However, models ultimately serve real-world, and it makes sense to extend data imputation techniques to practical scenarios. Firstly, in the analysis of historical traffic data and the development of traffic digital twin systems, it is essential to handle large volumes of sensor data and historical records. Research can focus on developing large-scale data imputation techniques to provide complete and accurate data support [53]. Secondly, autonomous driving systems, intelligent traffic signal systems, and traffic anomaly detection require real-time sensing of the surrounding environment and adaptation to dynamic traffic flows [37], [192]. Therefore, exploring real-time imputation techniques is crucial to ensure system continuity and efficient operation. Furthermore, during natural disasters or emergencies (such as earthquakes, floods, or power outages), as well as in remote areas with inadequate sensor and communication infrastructure coverage, the collected traffic data often have a large number of missing [70]. Future researches should enhance the ability of imputation techniques to recover from extreme data loss scenarios, reconstructing missing parts from limited data to provide as complete and robust traffic information support as possible. Inspired by these potential application scenarios,

researchers will face abundant research opportunities to drive the development and optimization of ITS.

## V. CONCLUSION

The absence of traffic spatio-temporal data values is a typical occurrence in ITS. Many excellent solutions have been put forth to address this issue. Therefore, we discuss the development of traffic data imputation to help future researchers interested in this topic get a quick overview of the field. In this paper, we first examined the potential causes of the missing traffic data and explained the impact of the missing as well as the necessity of imputation. Second, we incorporated the missing pattern of traffic data since the missing caused by different scenarios has a variety of manifestations and characteristics, which can affect the performance of the repair model. Thirdly, we summarized some commonly used public traffic datasets and model performance evaluation indicators in recent study, and also gave the data acquisition address and discuss their suitability for various types of research to make it easier for researchers to select the most appropriate dataset. Then, we systematically divided techniques into interpolation, statistical learning, and prediction methods based on the inherent characteristics of imputation methods. Finally, we delved into the primary open challenges in traffic data imputation, such as missing patterns, joint imputation, signal control effects, model robustness, and real-time applicability. Additionally, we offered humble insights into potential future development directions. We hope that this survey will prove useful to upcoming researchers.

## REFERENCES

- [1] D. Li, J. Cao, and Y. Yao, "Big data in smart cities," *Sci. China Inf. Sci.*, vol. 58, no. 10, pp. 1–12, Oct. 2015.
- [2] Y. Wang, C. An, J. Ou, Z. Lu, and J. Xia, "A general dynamic sequential learning framework for vehicle trajectory reconstruction using automatic vehicle location or identification data," *Phys. A, Stat. Mech. Appl.*, vol. 608, Dec. 2022, Art. no. 128243.
- [3] S. Raicu, M. Popa, and D. Costescu, "Uncertainties influencing transportation system performances," *Sustainability*, vol. 14, no. 13, p. 7660, Jun. 2022.
- [4] L. Li, Y. Li, and Z. Li, "Efficient missing data imputing for traffic flow by considering temporal and spatial dependence," *Transp. Res. C, Emerg. Technol.*, vol. 34, pp. 108–120, Sep. 2013.
- [5] H. Tan, G. Feng, J. Feng, W. Wang, Y.-J. Zhang, and F. Li, "A tensor-based method for missing traffic data completion," *Transp. Res. C, Emerg. Technol.*, vol. 28, pp. 15–27, Mar. 2013.
- [6] J. Yang, A.-O. Purevjav, and S. Li, "The marginal cost of traffic congestion and road pricing: Evidence from a natural experiment in Beijing," *Amer. Econ. J., Econ. Policy*, vol. 12, no. 1, pp. 418–453, Feb. 2020.
- [7] F. Tao, H. Zhang, A. Liu, and A. Y. C. Nee, "Digital twin in industry: State-of-the-art," *IEEE Trans. Ind. Informat.*, vol. 15, no. 4, pp. 2405–2415, Apr. 2019.
- [8] D. Omeiza, H. Webb, M. Jirotko, and L. Kunze, "Explanations in autonomous driving: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 10142–10162, Aug. 2022.
- [9] X. Kong, Z. Shen, K. Wang, G. Shen, and Y. Fu, "Exploring bus stop mobility pattern: A multi-pattern deep learning prediction framework," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 7, pp. 6604–6616, Jul. 2024.
- [10] X. Kong, H. Lin, R. Jiang, and G. Shen, "Anomalous sub-trajectory detection with graph contrastive self-supervised learning," *IEEE Trans. Veh. Technol.*, vol. 73, no. 7, pp. 9800–9811, Jul. 2024.
- [11] X. Kong, Z. Xu, G. Shen, J. Wang, Q. Yang, and B. Zhang, "Urban traffic congestion estimation and prediction based on floating car trajectory data," *Future Gener. Comput. Syst.*, vol. 61, pp. 97–107, Aug. 2016.

- [12] L. Deng, X.-Y. Liu, H. Zheng, X. Feng, and Y. Chen, "Graph spectral regularized tensor completion for traffic data imputation," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 10996–11010, Aug. 2022.
- [13] C. Chen, Y. Wang, L. Li, J. Hu, and Z. Zhang, "The retrieval of intraday trend and its influence on traffic prediction," *Transp. Res. C, Emerg. Technol.*, vol. 22, pp. 103–118, Jun. 2012.
- [14] Z. Zhang, X. Lin, M. Li, and Y. Wang, "A customized deep learning approach to integrate network-scale online traffic data imputation and prediction," *Transp. Res. C, Emerg. Technol.*, vol. 132, Nov. 2021, Art. no. 103372.
- [15] A. Wang, Y. Ye, X. Song, S. Zhang, and J. J. Q. Yu, "Traffic prediction with missing data: A multi-task learning approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 4, pp. 4189–4202, Apr. 2023.
- [16] F. M. Shrive, H. Stuart, H. Quan, and W. A. Ghali, "Dealing with missing data in a multi-question depression scale: A comparison of imputation methods," *BMC Med. Res. Methodol.*, vol. 6, no. 1, pp. 1–10, Dec. 2006.
- [17] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.
- [18] A. N. Baraldi and C. K. Enders, "An introduction to modern missing data analyses," *J. School Psychol.*, vol. 48, no. 1, pp. 5–37, Feb. 2010.
- [19] W. Zheng et al., "Integrating the traffic science with representation learning for city-wide network congestion prediction," *Inf. Fusion*, vol. 99, Nov. 2023, Art. no. 101837.
- [20] J. Li, L. Xu, R. Li, P. Wu, and Z. Huang, "Deep spatial-temporal bi-directional residual optimisation based on tensor decomposition for traffic data imputation on urban road network," *Appl. Intell.*, vol. 52, no. 10, pp. 11363–11381, Aug. 2022.
- [21] W.-C. Lin and C.-F. Tsai, "Missing value imputation: A review and analysis of the literature (2006–2017)," *Artif. Intell. Rev.*, vol. 53, no. 2, pp. 1487–1509, Feb. 2020.
- [22] T. Sun, S. Zhu, R. Hao, B. Sun, and J. Xie, "Traffic missing data imputation: A selective overview of temporal theories and algorithms," *Mathematics*, vol. 10, no. 14, p. 2544, Jul. 2022.
- [23] M. A. Shafique, "Imputing missing data in hourly traffic counts," *Sensors*, vol. 22, no. 24, p. 9876, Dec. 2022.
- [24] Y. Li, Z. Li, and L. Li, "Missing traffic data: Comparison of imputation methods," *IET Intell. Transp. Syst.*, vol. 8, no. 1, pp. 51–57, Feb. 2014.
- [25] R. K. C. Chan, J. M. Lim, and R. Parthiban, "Missing traffic data imputation for artificial intelligence in intelligent transportation systems: Review of methods, limitations, and challenges," *IEEE Access*, vol. 11, pp. 34080–34093, 2023.
- [26] H. V. Jagadish et al., "Big data and its technical challenges," *Commun. ACM*, vol. 57, no. 7, pp. 86–94, Jul. 2014.
- [27] M. Lei, A. Labbe, Y. Wu, and L. Sun, "Bayesian kernelized matrix factorization for spatiotemporal traffic data imputation and Kriging," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 18962–18974, Oct. 2022.
- [28] M. Liu, H. Huang, H. Feng, L. Sun, B. Du, and Y. Fu, "PriSTI: A conditional diffusion framework for spatiotemporal imputation," in *Proc. IEEE 39th Int. Conf. Data Eng. (ICDE)*, Apr. 2023, pp. 1927–1939.
- [29] J. L. Alcaraz and N. Strodthoff, "Diffusion-based time series imputation and forecasting with structured state space models," *Trans. Mach. Learn. Res.*, Feb. 2023. [Online]. Available: <https://openreview.net/forum?id=hHilbk7ApW>
- [30] W. Liang et al., "Spatial-temporal aware inductive graph neural network for C-ITS data recovery," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 8, pp. 1–12, Aug. 2022.
- [31] Y. Zhang, X. Wei, X. Zhang, Y. Hu, and B. Yin, "Self-attention graph convolution residual network for traffic data completion," *IEEE Trans. Big Data*, vol. 9, no. 2, pp. 528–541, Apr. 2023.
- [32] Y. Wang et al., "Attention-based message passing and dynamic graph convolution for spatiotemporal data imputation," *Sci. Rep.*, vol. 13, no. 1, p. 6887, Apr. 2023.
- [33] T. Nie, G. Qin, W. Ma, Y. Mei, and J. Sun, "ImputeFormer: Low rankness-induced transformers for generalizable spatiotemporal imputation," in *Proc. 30th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, vol. 44, Aug. 2024, pp. 2260–2271.
- [34] Z. Wang et al., "ST-GIN: An uncertainty quantification approach in traffic data imputation with spatio-temporal graph attention and bidirectional recurrent united neural networks," in *Proc. IEEE 26th Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2023, pp. 1454–1459.
- [35] Y. Liang, Z. Zhao, and L. Sun, "Memory-augmented dynamic graph convolution networks for traffic data imputation with diverse missing patterns," *Transp. Res. C, Emerg. Technol.*, vol. 143, Oct. 2022, Art. no. 103826.
- [36] A. Cini et al., "Filling the Gaps: Multivariate time series imputation by graph neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–20.
- [37] J.-M. Yang, Z.-R. Peng, and L. Lin, "Real-time spatiotemporal prediction and imputation of traffic status based on LSTM and graph Laplacian regularized matrix factorization," *Transp. Res. C, Emerg. Technol.*, vol. 129, Aug. 2021, Art. no. 103228.
- [38] C. Chen, "Freeway performance measurement system (PeMS)," Univ. California, Berkeley, CA, USA, Tech. Rep. AAI3082138, 2002.
- [39] Z. Cai, Y. Shu, X. Su, L. Guo, and Z. Ding, "A traffic data interpolation method for IoT sensors based on spatio-temporal dependence," *Internet Things*, vol. 21, Apr. 2023, Art. no. 100648.
- [40] B. Sun, L. Ma, W. Cheng, W. Wen, P. Goswami, and G. Bai, "An improved k-nearest neighbours method for traffic time series imputation," in *Proc. Chin. Autom. Congr. (CAC)*, Oct. 2017, pp. 7346–7351.
- [41] Y. Chen, Y. Lv, and F.-Y. Wang, "Traffic flow imputation using parallel data and generative adversarial networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 4, pp. 1624–1630, Apr. 2020.
- [42] P. Sure, C. P. Srinivasan, and C. N. Babu, "Spatio-temporal constraint-based low rank matrix completion approaches for road traffic networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 13452–13462, Aug. 2022.
- [43] N. Zhao, Z. Li, and Y. Li, "Improving the traffic data imputation accuracy using temporal and spatial information," in *Proc. 7th Int. Conf. Intell. Comput. Technol. Autom.*, Oct. 2014, pp. 312–317.
- [44] L. Han, K. Zheng, L. Zhao, X. Wang, and H. Wen, "Content-aware traffic data completion in ITS based on generative adversarial nets," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 11950–11962, Oct. 2020.
- [45] L. Hu, Y. Jia, W. Chen, L. Wen, and Z. Ye, "A flexible and robust tensor completion approach for traffic data recovery with low-rankness," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 3, pp. 2558–2572, Mar. 2024.
- [46] B. Zhang, R. Miao, and Z. Chen, "Spatial-temporal traffic data imputation based on dynamic multi-level generative adversarial networks for urban governance," *Appl. Soft Comput.*, vol. 151, Jan. 2024, Art. no. 111128.
- [47] X. Song, Y. Ye, and J. J. Yu, "TINet: Multi-dimensional traffic data imputation via transformer network," in *Proc. Int. Conf. Artif. Neural Netw.* Cham, Switzerland: Springer, 2021, pp. 306–317.
- [48] L. Li, J. Zhang, Y. Wang, and B. Ran, "Missing value imputation for traffic-related time series data based on a multi-view learning method," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 8, pp. 2933–2943, Aug. 2019.
- [49] C. Qiu, Y. Li, M. Kang, D. Chen, and W. Yu, "CDSTTN: A data imputation method for cyber-physical systems by causal dense spatial-temporal transformer network," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 13, no. 3, pp. 851–860, May 2023.
- [50] W. Zhong, Q. Suo, X. Jia, A. Zhang, and L. Su, "Heterogeneous spatio-temporal graph convolution network for traffic forecasting with missing values," in *Proc. IEEE 41st Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Jul. 2021, pp. 707–717.
- [51] Y. Chen and X. Chen, "A novel reinforced dynamic graph convolutional network model with data imputation for network-wide traffic flow prediction," *Transp. Res. C, Emerg. Technol.*, vol. 143, Oct. 2022, Art. no. 103820.
- [52] D. Xu, H. Peng, Y. Tang, and H. Guo, "Hierarchical spatio-temporal graph convolutional neural networks for traffic data imputation," *Inf. Fusion*, vol. 106, Jun. 2024, Art. no. 102292.
- [53] K. Zhang, F. Zhou, L. Wu, N. Xie, and Z. He, "Semantic understanding and prompt engineering for large-scale traffic data imputation," *Inf. Fusion*, vol. 102, Feb. 2024, Art. no. 102038.
- [54] G. Boquet, A. Morell, J. Serrano, and J. L. Vicario, "A variational autoencoder solution for road traffic forecasting systems: Missing data imputation, dimension reduction, model selection and anomaly detection," *Transp. Res. C, Emerg. Technol.*, vol. 115, Jun. 2020, Art. no. 102622.
- [55] Y. Duan, Y. Lv, Y.-L. Liu, and F.-Y. Wang, "An efficient realization of deep learning for traffic data imputation," *Transp. Res. C, Emerg. Technol.*, vol. 72, pp. 168–181, Nov. 2016.
- [56] Z. Zeng, B. Liu, J. Feng, and X. Yang, "Low-rank tensor and hybrid smoothness regularization-based approach for traffic data imputation with multimodal missing," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 10, pp. 13014–13026, Oct. 2024.
- [57] X. Chen, Y. Chen, N. Saunier, and L. Sun, "Scalable low-rank tensor learning for spatiotemporal traffic data imputation," *Transp. Res. C, Emerg. Technol.*, vol. 129, Aug. 2021, Art. no. 103226.



- [58] T. Nie, G. Qin, Y. Wang, and J. Sun, "Correlating sparse sensing for large-scale traffic speed estimation: A laplacian-enhanced low-rank tensor Kriging approach," *Transp. Res. C, Emerg. Technol.*, vol. 152, Jul. 2023, Art. no. 104190.
- [59] G. Shen, N. Liu, Y. Liu, W. Zhou, and X. Kong, "Traffic flow imputation based on multi-perspective spatiotemporal generative adversarial networks," in *Proc. CECNet*, 2022, pp. 62–73.
- [60] X. Kong, W. Zhou, G. Shen, W. Zhang, N. Liu, and Y. Yang, "Dynamic graph convolutional recurrent imputation network for spatiotemporal traffic missing data," *Knowl.-Based Syst.*, vol. 261, Feb. 2023, Art. no. 110188.
- [61] G. Shen, W. Zhou, W. Zhang, N. Liu, Z. Liu, and X. Kong, "Bidirectional spatial-temporal traffic data imputation via graph attention recurrent neural network," *Neurocomputing*, vol. 531, pp. 151–162, Apr. 2023.
- [62] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting," in *Proc. Int. Conf. Learn. Represent.*, 2018.
- [63] X. Wang et al., "An observed value consistent diffusion model for imputing missing values in multivariate time series," in *Proc. 29th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2023, pp. 2409–2418.
- [64] P. Wu, M. Ding, and Y. Zheng, "Spatiotemporal traffic data imputation by synergizing low tensor ring rank and nonlocal subspace regularization," *IET Intell. Transp. Syst.*, vol. 17, no. 9, pp. 1908–1923, Sep. 2023.
- [65] Z. Cui, R. Ke, Z. Pu, and Y. Wang, "Stacked bidirectional and unidirectional LSTM recurrent neural network for forecasting network-wide traffic state with missing values," *Transp. Res. C, Emerg. Technol.*, vol. 118, Sep. 2020, Art. no. 102674.
- [66] Z. Cui, K. Henrickson, R. Ke, and Y. Wang, "Traffic graph convolutional recurrent neural network: A deep learning framework for network-scale traffic learning and forecasting," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 11, pp. 4883–4894, Nov. 2020.
- [67] X. Chen, J. Yang, and L. Sun, "A nonconvex low-rank tensor completion model for spatiotemporal traffic data imputation," *Transp. Res. C, Emerg. Technol.*, vol. 117, Aug. 2020, Art. no. 102673.
- [68] H. Li, Q. Cao, Q. Bai, Z. Li, and H. Hu, "Multistate time series imputation using generative adversarial network with applications to traffic data," *Neural Comput. Appl.*, vol. 35, no. 9, pp. 6545–6567, Mar. 2023.
- [69] W. Zhang, P. Zhang, Y. Yu, X. Li, S. A. Biancardo, and J. Zhang, "Missing data repairs for traffic flow with self-attention generative adversarial imputation net," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 7919–7930, May 2021.
- [70] X. Chen, M. Lei, N. Saunier, and L. Sun, "Low-rank autoregressive tensor completion for spatiotemporal traffic data imputation," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 12301–12310, Aug. 2021.
- [71] X. Xu, M. Lin, X. Luo, and Z. Xu, "HRST-LR: A Hessian regularization spatiotemporal low rank algorithm for traffic data imputation," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 10, pp. 1–17, Oct. 2023.
- [72] D. Xu, Z. Yu, T. Tian, and Y. Yang, "Generative adversarial network for imputation of road network traffic state data," in *Proc. China Nat. Conf. Big Data Social Comput.* Cham, Switzerland: Springer, 2022, pp. 80–96.
- [73] H. Chen, M. Lin, J. Liu, H. Wang, C. Zhang, and Z. Xu, "NT-DPTC: A non-negative temporal dimension preserved tensor completion model for missing traffic data imputation," *Inf. Sci.*, vol. 653, Jan. 2024, Art. no. 119797.
- [74] B.-Z. Li, X.-L. Zhao, X. Chen, M. Ding, and R. W. Liu, "Convolutional low-rank tensor representation for structural missing traffic data imputation," *IEEE Trans. Intell. Transp. Syst.*, early access, Jul. 31, 2024, doi: [10.1109/TITS.2024.3430039](https://doi.org/10.1109/TITS.2024.3430039).
- [75] Y. He, Y. Jia, Y. Jia, C. An, Z. Lu, and J. Xia, "An integrated intra-view and inter-view framework for multiple traffic variable data simultaneous recovery," *IEEE Trans. Intell. Transp. Syst.*, early access, Jun. 20, 2024, doi: [10.1109/TITS.2024.3414506](https://doi.org/10.1109/TITS.2024.3414506).
- [76] T. Nie, G. Qin, and J. Sun, "Truncated tensor Schatten p-norm based approach for spatiotemporal traffic data imputation with complicated missing patterns," *Transp. Res. C, Emerg. Technol.*, vol. 141, Aug. 2022, Art. no. 103737.
- [77] W. Yue, D. Zhou, S. Wang, and P. Duan, "Engineering traffic prediction with online data imputation: A graph-theoretic perspective," *IEEE Syst. J.*, vol. 17, no. 3, pp. 1–12, May 2023.
- [78] X. Chen, Z. He, and J. Wang, "Spatial-temporal traffic speed patterns discovery and incomplete data recovery via SVD-combined tensor decomposition," *Transp. Res. C, Emerg. Technol.*, vol. 86, pp. 59–77, Jan. 2018.
- [79] X. Chen, Z. He, and L. Sun, "A Bayesian tensor decomposition approach for spatiotemporal traffic data imputation," *Transp. Res. C, Emerg. Technol.*, vol. 98, pp. 73–84, Jan. 2019.
- [80] T. Zhang, D.-G. Zhang, H.-R. Yan, J.-N. Qiu, and J.-X. Gao, "A new method of data missing estimation with FNN-based tensor heterogeneous ensemble learning for Internet of Vehicle," *Neurocomputing*, vol. 420, pp. 98–110, Jan. 2021.
- [81] C. Lyu, Q.-L. Lu, X. Wu, and C. Antoniou, "Tucker factorization-based tensor completion for robust traffic data imputation," *Transp. Res. C, Emerg. Technol.*, vol. 160, Mar. 2024, Art. no. 104502.
- [82] X. Chen, Z. He, Y. Chen, Y. Lu, and J. Wang, "Missing traffic data imputation and pattern discovery with a Bayesian augmented tensor factorization model," *Transp. Res. C, Emerg. Technol.*, vol. 104, pp. 66–77, Jul. 2019.
- [83] J. A. Deri and J. M. F. Moura, "Taxi data in New York City: A network perspective," in *Proc. 49th Asilomar Conf. Signals, Syst. Comput.*, Nov. 2015, pp. 1829–1833.
- [84] B. Yang, Y. Kang, Y. Yuan, X. Huang, and H. Li, "ST-LBAGAN: Spatio-temporal learnable bidirectional attention generative adversarial networks for missing traffic data imputation," *Knowl.-Based Syst.*, vol. 215, Mar. 2021, Art. no. 106705.
- [85] J. Zhang, Y. Zheng, and D. Qi, "Deep spatio-temporal residual networks for citywide crowd flows prediction," in *Proc. AAAI Conf. Artif. Intell.*, 2017, vol. 31, no. 1.
- [86] S. Zhang, X. Hu, W. Zhang, J. Chen, and H. Huang, "Learning traffic as videos: A spatio-temporal VAE approach to periodic traffic raster data imputation," *Intell. Data Anal.*, pp. 1–22, Feb. 2024.
- [87] J. Chen, S. Zhang, W. Zhang, J. Chen, C. Gu, and H. Huang, "MTSVAE: A traffic data imputation model considering different periodic temporal and global spatial features," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2022, pp. 1–8.
- [88] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "Driving with knowledge from the physical world," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2011, pp. 316–324.
- [89] A. Ben Said and A. Erradi, "Spatiotemporal tensor completion for improved urban traffic imputation," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 6836–6849, Jul. 2022.
- [90] P. D. Allison, *Missing Data*. Newbury Park, CA, USA: Sage, 2001.
- [91] R. J. Little and D. B. Rubin, *Statistical Analysis With Missing Data*, vol. 793. Hoboken, NJ, USA: Wiley, 2019.
- [92] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*, vol. 81. Hoboken, NJ, USA: Wiley, 2004.
- [93] P. Wu, L. Xu, and Z. Huang, "Imputation methods used in missing traffic data: A literature review," in *Proc. ISICA*, Guangzhou, China. Cham, Switzerland: Springer, 2019, pp. 662–677.
- [94] T. Zhang, J. Wang, and J. Liu, "A gated generative adversarial imputation approach for signalized road networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 12144–12160, Aug. 2022.
- [95] B. Bae, H. Kim, H. Lim, Y. Liu, L. D. Han, and P. B. Freeze, "Missing data imputation for traffic flow speed using spatio-temporal cokriging," *Transp. Res. C, Emerg. Technol.*, vol. 88, pp. 124–139, Mar. 2018.
- [96] Z. Wang, R. Chu, M. Zhang, X. Wang, and S. Luan, "An improved selective ensemble learning method for highway traffic flow state identification," *IEEE Access*, vol. 8, pp. 212623–212634, 2020.
- [97] J. Tang, G. Zhang, Y. Wang, H. Wang, and F. Liu, "A hybrid approach to integrate fuzzy C-means based imputation method with genetic algorithm for missing traffic volume data estimation," *Transp. Res. C, Emerg. Technol.*, vol. 51, pp. 29–40, Feb. 2015.
- [98] Q. Shang, Z. Yang, S. Gao, and D. Tan, "An imputation method for missing traffic data based on FCM optimized by PSO-SVR," *J. Adv. Transp.*, vol. 2018, pp. 1–21, Jan. 2018.
- [99] S. Tak, S. Woo, and H. Yeo, "Data-driven imputation method for traffic data in sectional units of road links," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 6, pp. 1762–1771, Jun. 2016.
- [100] D. Adhikari et al., "A comprehensive survey on imputation of missing data in Internet of Things," *ACM Comput. Surv.*, vol. 55, no. 7, pp. 1–38, Dec. 2022.
- [101] X. Wu et al., "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1–37, 2008.
- [102] G. E. Batista and M. C. Monard, "A study of K-nearest neighbour as an imputation method," *HIS*, vol. 87, nos. 251–260, p. 48, Dec. 2002.
- [103] H. Yang et al., "A Kriging based spatiotemporal approach for traffic volume data imputation," *PLoS ONE*, vol. 13, no. 4, Apr. 2018, Art. no. e0195957.

- [104] X. Ma, S. Luan, C. Ding, H. Liu, and Y. Wang, "Spatial interpolation of missing annual average daily traffic data using copula-based model," *IEEE Intell. Transp. Syst. Mag.*, vol. 11, no. 3, pp. 158–170, Fall 2019.
- [105] D. G. Krige, "A statistical approach to some basic mine valuation problems on the Witwatersrand," *J. Southern Afr. Inst. Mining Metall.*, vol. 52, no. 6, pp. 119–139, 1951.
- [106] D. Marcotte, "Cokriging with MATLAB," *Comput. Geosci.*, vol. 17, no. 9, pp. 1265–1280, 1991.
- [107] Y. Wu, D. Zhuang, A. Labbe, and L. Sun, "Inductive graph neural networks for spatiotemporal kriging," in *Proc. AAAI Conf. Artif. Intell.*, May 2021, vol. 35, no. 5, pp. 4478–4485.
- [108] S. Thajchayapong and J. A. Barria, "Spatial inference of traffic transition using micro-macro traffic variables," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 854–864, Apr. 2015.
- [109] T. Nie, G. Qin, Y. Wang, and J. Sun, "Towards better traffic volume estimation: Jointly addressing the underdetermination and nonequilibrium problems with correlation-adaptive GNNs," *Transp. Res. C, Emerg. Technol.*, vol. 157, Dec. 2023, Art. no. 104402.
- [110] X. Miao, Y. Wu, L. Chen, Y. Gao, and J. Yin, "An experimental survey of missing data imputation algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 7, pp. 6630–6650, Jun. 2023.
- [111] L. Qu, J. Hu, L. Li, and Y. Zhang, "PPCA-based missing data imputation for traffic flow volume: A systematical approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 3, pp. 512–522, Sep. 2009.
- [112] Y. Li, Z. Li, L. Li, Y. Zhang, and M. Jin, "Comparison on PPCA, KPPCA and MPPCA based missing data imputing for traffic flow," in *Proc. ICTIS*, vol. 39, Jun. 2013, pp. 1151–1156.
- [113] A. Liu, C. Li, W. Yue, and X. Zhou, "Real-time traffic prediction: A novel imputation optimization algorithm with missing data," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–7.
- [114] B. Agrawal, T. Wiktoriski, and C. Rong, "Adaptive anomaly detection in cloud using robust and scalable principal component analysis," in *Proc. 15th Int. Symp. Parallel Distrib. Comput. (ISPDC)*, Jul. 2016, pp. 100–106.
- [115] Y. Wang, Y. Zhang, X. Piao, H. Liu, and K. Zhang, "Traffic data reconstruction via adaptive spatial-temporal correlations," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 4, pp. 1531–1543, Apr. 2019.
- [116] Z. Long, Y. Liu, L. Chen, and C. Zhu, "Low rank tensor completion for multiway visual data," *Signal Process.*, vol. 155, pp. 301–316, Feb. 2019.
- [117] Y. Zhu, W. Wang, G. Yu, J. Wang, and L. Tang, "A Bayesian robust CP decomposition approach for missing traffic data imputation," *Multimedia Tools Appl.*, vol. 81, no. 23, pp. 33171–33184, Apr. 2022.
- [118] Y. Zhu, J. Wang, J. Wang, and Z. He, "Multitask neural tensor factorization for road traffic speed-volume correlation pattern learning and joint imputation," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 24550–24560, Dec. 2022.
- [119] J. Xing, R. Liu, K. Anish, and Z. Liu, "A customized data fusion tensor approach for interval-wise missing network volume imputation," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 11, pp. 12107–12122, Nov. 2023.
- [120] J. Yoon, J. Jordon, and M. Schaar, "GAIN: Missing data imputation using generative adversarial nets," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5689–5698.
- [121] A. Kazemi and H. Meidani, "IGANI: Iterative generative adversarial networks for imputation with application to traffic data," *IEEE Access*, vol. 9, pp. 112966–112977, 2021.
- [122] Y. Yuan, Y. Zhang, B. Wang, Y. Peng, Y. Hu, and B. Yin, "STGAN: Spatio-temporal generative adversarial network for traffic data imputation," *IEEE Trans. Big Data*, vol. 9, no. 1, pp. 200–211, Feb. 2023.
- [123] N. Gao et al., "Generative adversarial networks for spatio-temporal data: A survey," *ACM Trans. Intell. Syst. Technol.*, vol. 13, no. 2, pp. 1–25, Apr. 2022.
- [124] Y. Ye, S. Zhang, and J. J. Yu, "Spatial-temporal traffic data imputation via graph attention convolutional network," in *Proc. Int. Conf. Artif. Neural Netw. (ICANN)*. Cham, Switzerland: Springer, 2021, pp. 241–252.
- [125] R. Wang, M. Li, Q. Guo, Y. Xiao, and Z. Yang, "Road network pixelization: A traffic flow imputation method based on image restoration techniques," *Expert Syst. Appl.*, vol. 237, Mar. 2024, Art. no. 121468.
- [126] L. Qu, Y. Zhang, J. Hu, L. Jia, and L. Li, "A BPCA based missing value imputing method for traffic flow volume data," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2008, pp. 985–990.
- [127] G. Pastor, "A low-rank tensor model for imputation of missing vehicular traffic volume," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8934–8938, Sep. 2018.
- [128] Y. Tashiro, J. Song, Y. Song, and S. Ermon, "CSDI: Conditional score-based diffusion models for probabilistic time series imputation," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2021, pp. 24804–24816.
- [129] S. Zhang, S. Wang, X. Tan, R. Liu, J. Zhang, and J. Wang, "Sasdim: Self-adaptive noise scaling diffusion model for spatial time series imputation," 2023, *arXiv:2309.01988*.
- [130] Y. M. Hwang, S.-C. Son, N. Kim, S.-K. Ko, and B.-T. Lee, "RDMI: Recursive training-based diffusion model for multivariate time series imputation," in *Proc. Int. Tech. Conf. Circuits/Syst., Comput., Commun. (ITC-CSCC)*, vol. 31, Jun. 2023, pp. 1–5.
- [131] L. Yang et al., "Diffusion models: A comprehensive survey of methods and applications," *ACM Comput. Surv.*, vol. 56, no. 4, pp. 1–39, 2023.
- [132] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. van Gool, "RePaint: Inpainting using denoising diffusion probabilistic models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11461–11471.
- [133] Z. C. Lipton, "The Mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.
- [134] M. Zhong, S. Sharma, and P. Lingras, "Genetically designed models for accurate imputation of missing traffic counts," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1879, no. 1, pp. 71–79, Jan. 2004.
- [135] M. Elshenawy, M. El-darieby, and B. Abdulhai, "Automatic imputation of missing highway traffic volume data," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops (PerCom Workshops)*, Mar. 2018, pp. 373–378.
- [136] B. Ghosh, B. Basu, and M. O'Mahony, "Time-series modelling for forecasting vehicular traffic flow in Dublin," in *Proc. 84th Annu. Meeting Transp. Res. Board*, Washington, DC, USA, 2005.
- [137] M. Zhong, S. Sharma, and Z. Liu, "Assessing robustness of imputation models based on data from different jurisdictions: Examples of Alberta and Saskatchewan, Canada," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1917, pp. 116–126, Jan. 2005.
- [138] J. E. Contreras-Reyes, "Rényi entropy and divergence for VARFIMA processes based on characteristic and impulse response functions," *Chaos, Solitons Fractals*, vol. 160, Jul. 2022, Art. no. 112268.
- [139] C. Fang and C. Wang, "Time series data imputation: A survey on deep learning approaches," 2020, *arXiv:2011.11347*.
- [140] I. B. Aydılek and A. Arslan, "A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm," *Inf. Sci.*, vol. 233, pp. 25–35, Jun. 2013.
- [141] X. Luo, X. Meng, W. Gan, and Y. Chen, "Traffic data imputation algorithm based on improved low-rank matrix decomposition," *J. Sensors*, vol. 2019, pp. 1–11, Jul. 2019.
- [142] Y.-Y. Choi, H. Shon, Y.-J. Byon, D.-K. Kim, and S. Kang, "Enhanced application of principal component analysis in machine learning for imputation of missing traffic data," *Appl. Sci.*, vol. 9, no. 10, p. 2149, May 2019.
- [143] L. Li, J. Zhang, F. Yang, and B. Ran, "Robust and flexible strategy for missing data imputation in intelligent transportation system," *IET Intell. Transp. Syst.*, vol. 12, no. 2, pp. 151–157, Mar. 2018.
- [144] L. R. Medsker and L. Jain, "Recurrent neural networks," *Des. Appl.*, vol. 5, pp. 64–67, Dec. 2001.
- [145] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [146] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*.
- [147] Y. Tian, K. Zhang, J. Li, X. Lin, and B. Yang, "LSTM-based traffic flow prediction with missing data," *Neurocomputing*, vol. 318, pp. 297–305, Nov. 2018.
- [148] W. Cao, D. Wang, J. Li, H. Zhou, L. Li, and Y. Li, "BRITS: Bidirectional recurrent imputation for time series," in *Advances in Neural Information Processing Systems*, vol. 31. Red Hook, NY, USA: Curran Associates, 2018.
- [149] V. A. Le, T. T. Le, P. L. Nguyen, H. T. T. Binh, R. Akerkar, and Y. Ji, "GCRINT: Network traffic imputation using graph convolutional recurrent neural network," in *Proc. IEEE Int. Conf. Commun.*, Jul. 2021, pp. 1–6.
- [150] D. Zhu, G. Shen, J. Chen, W. Zhou, and X. Kong, "A higher-order motif-based spatiotemporal graph imputation approach for transportation networks," *Wireless Commun. Mobile Comput.*, vol. 2022, pp. 1–16, Jan. 2022.

- [151] J. Ming et al., "Multi-graph convolutional recurrent network for fine-grained lane-level traffic flow imputation," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2022, pp. 348–357.
- [152] J. Kwon, C. Cha, and H. Park, "Multilayered lstm with parameter transfer for vehicle speed data imputation," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Jun. 2021, pp. 1–5.
- [153] S. Zhang, C. Zhang, S. Zhang, and J. James, "Attention-driven recurrent imputation for traffic speed," *IEEE Open J. Intell. Transp. Syst.*, vol. 3, pp. 723–737, 2022.
- [154] Y. Luo, X. Cai, Y. Zhang, J. Xu, and Y. Xiaojie, "Multivariate time series imputation with generative adversarial networks," in *Advances in Neural Information Processing Systems*, vol. 31. Red Hook, NY, USA: Curran Associates, 2018.
- [155] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Sci. Rep.*, vol. 8, no. 1, p. 6085, Apr. 2018.
- [156] Z. Ji and W. Zhu, "A traffic data imputing method based on multi-source recurrent neural network," *Proc. SPIE*, vol. 12260, pp. 90–95, Aug. 2022.
- [157] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1310–1318.
- [158] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30. Red Hook, NY, USA: Curran Associates, 2017.
- [159] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, Mar. 2020.
- [160] S. Guo, Y. Lin, N. Feng, C. Song, and H. Wan, "Attention based spatial-temporal graph convolutional networks for traffic flow forecasting," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 922–929.
- [161] X. Geng et al., "Spatiotemporal multi-graph convolution network for ride-hailing demand forecasting," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 3656–3663.
- [162] Q. Zhang, J. Chang, G. Meng, S. Xiang, and C. Pan, "Spatio-temporal graph structure learning for traffic forecasting," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 1, pp. 1177–1185.
- [163] W. Chen, L. Chen, Y. Xie, W. Cao, and X. Feng, "Multi-range attentive bicomponent graph convolutional network for traffic forecasting," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 4, 2020, pp. 3529–3536.
- [164] W. Jiang and J. Luo, "Graph neural network for traffic forecasting: A survey," *Expert Syst. Appl.*, vol. 207, Nov. 2022, Art. no. 117921.
- [165] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.
- [166] R. Liu, Y. Kan, S. Zhao, B. Cheng, Z. Ma, and W. Wu, "Turning traffic volume imputation for persistent missing patterns with GNNs," *Appl. Intell.*, vol. 53, no. 1, pp. 491–508, Jan. 2023.
- [167] C. Zhao, A. Song, Y. Du, and B. Yang, "TrajGAT: A map-embedded graph attention network for real-time vehicle trajectory imputation of roadside perception," *Transp. Res. C, Emerg. Technol.*, vol. 142, Sep. 2022, Art. no. 103787.
- [168] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," 2017, *arXiv:1710.10903*.
- [169] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 3634–3640.
- [170] H. Zeng et al., "Decoupling the depth and scope of graph neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 19665–19679.
- [171] A. Loukas, "What graph neural networks cannot learn: Depth vs width," in *Proc. Int. Conf. Learn. Represent.*, 2020.
- [172] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
- [173] M. Chen et al., "Scalable graph neural networks via bidirectional propagation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 14556–14566.
- [174] W.-L. Chiang, X. Liu, S. Si, Y. Li, S. Bengio, and C.-J. Hsieh, "Cluster-GCN: An efficient algorithm for training deep and large graph convolutional networks," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 257–266.
- [175] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986.
- [176] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.
- [177] Y. Zhuang, R. Ke, and Y. Wang, "Innovative method for traffic data imputation based on convolutional neural network," *IET Intell. Transp. Syst.*, vol. 13, no. 4, pp. 605–613, Apr. 2019.
- [178] Y. Chen, K. Shi, X. Wang, and G. Xu, "MTSTI: A multi-task learning framework for spatiotemporal imputation," in *Proc. Int. Conf. Adv. Data Mining Appl.* Cham, Switzerland: Springer, 2023, pp. 180–194.
- [179] Y. Qu, Z. Li, X. Zhao, and J. Ou, "Towards real-world traffic prediction and data imputation: A multi-task pretraining and fine-tuning approach," *Inf. Sci.*, vol. 657, Feb. 2024, Art. no. 119972.
- [180] Y. Duan, Y. Lv, W. Kang, and Y. Zhao, "A deep learning based approach for traffic data imputation," in *Proc. 17th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2014, pp. 912–917.
- [181] Y. Ye, S. Zhang, and J. J. Q. Yu, "Traffic data imputation with ensemble convolutional autoencoder," in *Proc. IEEE Int. Intell. Transp. Syst. Conf. (ITSC)*, Sep. 2021, pp. 1340–1345.
- [182] G. Boquet, J. L. Vicario, A. Morell, and J. Serrano, "Missing data in traffic estimation: A variational autoencoder imputation method," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 2882–2886.
- [183] S. Zhang, X. Hu, J. Chen, W. Zhang, and H. Huang, "An effective variational auto-encoder-based model for traffic flow imputation," *Neural Comput. Appl.*, vol. 36, no. 5, pp. 2617–2631, Feb. 2024.
- [184] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.
- [185] I. Higgins et al., "beta-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. Int. Conf. Learn. Represent.*, 2016.
- [186] H. Kim and A. Mnih, "Disentangling by factorising," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2649–2658.
- [187] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers," in *Proc. 11th Int. Conf. Learn. Represent.*, 2022.
- [188] Y. Zhang, L. Ma, S. Pal, Y. Zhang, and M. Coates, "Multi-resolution time-series transformer for long-term forecasting," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2024, pp. 4222–4230.
- [189] X. Chen and Q. Li, "Event modeling and mining: A long journey toward explainable events," *VLDB J.*, vol. 29, no. 1, pp. 459–482, Jan. 2020.
- [190] C. Liu et al., "Spatial-temporal large language model for traffic prediction," 2024, *arXiv:2401.10134*.
- [191] M. Jin et al., "Time-LLM: Time series forecasting by reprogramming large language models," in *Proc. 12th Int. Conf. Learn. Represent.*, 2023.
- [192] Y. Tong et al., "The simpler the better: A unified approach to predicting original taxi demands based on large-scale online platforms," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2017, pp. 1653–1662.



**Yimei Zhang** received the B.S. degree in data science and big data technology from Wenzhou University, Wenzhou, China, in 2022. She is currently pursuing the Ph.D. degree in electronic information with Zhejiang University of Technology, Hangzhou, China. Her research interests include social computing, urban science, and data mining.



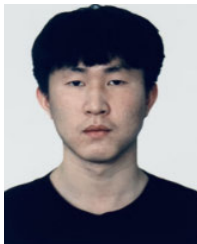
**Xiangjie Kong** (Senior Member, IEEE) received the B.Sc. and Ph.D. degrees from Zhejiang University, Hangzhou, China. He is currently a Full Professor with the College of Computer Science and Technology, Zhejiang University of Technology. Previously, he was an Associate Professor with the School of Software, Dalian University of Technology, China. He has published over 200 scientific papers in international journals and conferences (with over 170 indexed by ISI SCIE). His research interests include network science, mobile computing, and urban computing. He is a Senior Member of CCF and a member of ACM.



**Yanjie Fu** (Senior Member, IEEE) received the B.E. degree from the University of Science and Technology of China in 2008, the M.E. degree from Chinese Academy of Sciences, in 2011, and the Ph.D. degree from the Rutgers, The State University of New Jersey, in 2016. He is currently an Associate Professor with the School of Computing and AI, Arizona State University. He has research experience in industry research labs, such as Microsoft Research Asia and the IBM Thomas J. Watson Research Center. He was chosen for the US NAE FOE early career engineer in 2023. He is committed to data science education. His graduated Ph.D. students have joined academia as a tenure-track faculty members. He is broadly interested in data mining, machine learning, and their interdisciplinary applications. His research aims to develop robust machine intelligence with imperfect and complex data by building tools to address framework, algorithmic, data, and computing challenges. His recent focuses are spatial-temporal AI, graph learning, reinforcement learning, learning with unlabeled data, stream learning, and distribution drift. He is a Senior Member of ACM.



**Wenfeng Zhou** received the B.S. degree from the College of Science, Jiangxi University of Science and Technology, Jiangxi, China, in 2020. He is currently pursuing the Ph.D. degree in electronic information in computer science and technology with the School of Computer Science, Zhejiang University of Technology, China. His research interests include traffic data mining and intelligent transportation systems.



**Jin Liu** received the B.Sc. degree in Internet of Things engineering from Zhejiang University of Technology, Hangzhou, China, in 2022, where he is currently pursuing the master's degree. His research interests include social computing, urban science, and compute vision.



**Guojian Shen** received the B.Sc. degree in control theory and control engineering and the Ph.D. degree in control science and engineering from Zhejiang University, Hangzhou, China, in 1999 and 2004, respectively. He is currently a Professor with the College of Computer Science and Technology, Zhejiang University of Technology. His current research interests include artificial intelligence theory, big data analytics, and intelligent transportation systems.